Evaluation et enseignement supérieur : un couple maudit, au bord du divorce ?

Romainville, Marc

Published in:
Evaluation et enseignement supérieur

Publication date:
2013

Document Version
le PDF de l'éditeur

Chapitre conclusif
Évaluation et enseignement supérieur : un couple maudit, au bord du divorce ?

par Marc ROMAINVILLE*

Le présent chapitre conclusif comprend quatre parties. Dans la première, on rappellera d'abord combien l'évaluation est actuellement considérée comme suspecte aux yeux de beaucoup d'acteurs de l'enseignement supérieur. On cherchera ensuite, dans la deuxième partie, à identifier les raisons précises de la défiance dont l'évaluation fait aujourd'hui l'objet alors que, paradoxalement, les pratiques effectives d'évaluation font et ont toujours fait partie de l'exercice ordinaire des métiers de l'enseignement supérieur. La troisième partie sera consacrée à une esquisse de synthèse des chapitres précédents et plus généralement des travaux présentés à l'occasion du 23e colloque de l'ADMEE dont cet ouvrage est issu, en cherchant à identifier les principaux objets sur lesquels portent les recherches les plus actuelles en matière d'évaluation dans l'enseignement supérieur. On montrera enfin, dans la quatrième partie, que ces recherches mettent au jour une série de tensions transversales qui permettent de suggérer autant de rééquilibrages à opérer au sein des pratiques évaluatives. Un de ces équilibres à rétablir a trait à la nécessite de concilier, d'une part, le souci légitime de formaliser les pratiques d'évaluation au nom d'un certain nombre de principes peu contestables (tels que la transparence et la collégialité des décisions), et, d'autre part, les effets délétères et contre-productifs d'une formalisation excessive et en définitive illusoire. Dit plus crûment, trop d'évaluation formalisée finit par tuer l'évaluation et un équilibre délicat est à trouver entre, à une extrémité, une quasi absence d'explicitation des procédures et des normes qui conduit tout droit à l'arbitraire et renforce les relations de pouvoir et, à l'autre extrémité, une hyper-explicitation qui alourdit considérablement l'exercice des métiers du supérieur et qui, au nom d'une objectivité chimérique, masque les réels enjeux de l'évaluation et diffère le débat sur les valeurs qu'elle charrie inévitablement.

* Université de Namur.

Dans Évaluation et enseignement supérieur (dir. Marc Romainville, Rémi Goasdoué et Marc Vantourout), Bruxelles, De Boeck, 2013.

1. UNE DÉFIANCE TOUS AZIMUTS...

L'évaluation n'a, de nos jours, pas bonne presse et c'est un euphémisme dans l'enseignement supérieur. On ne compte plus les pamphlets sévères dénonçant la multiplication des procédures d'évaluation à tous les niveaux. On évaluerait tout (les personnes, leurs actions et leur *efficacité*[100]), n'importe quoi et surtout n'importe comment. Sont notamment dénoncées la douce folie de la mesure et la nouvelle manie du classement qui se seraient emparées de l'enseignement supérieur. Dans le domaine de l'évaluation institutionnelle (des laboratoires, départements et établissements ou de la *productivité* en recherche), certains y voient le signe d'une nouvelle idéologie[101], d'une grande imposture qui avancerait masquée : le recours à des méthodes d'évaluation de plus en plus sophistiquées nous ferait considérer comme objectifs et découlant de procédures neutres des choix authentiquement politiques. L'idéologie de l'évaluation consisterait donc à dissimuler l'exercice du pouvoir derrière des procédures qui se présentent comme froides, mécaniques et rigoureuses (Zarka, 2009).

Au quotidien, les professeurs de l'enseignement supérieur déplorent le prurit évaluatif dont ils seraient tout à la fois les victimes et les acteurs forcés : leur activité serait sans cesse évaluée, comme leur productivité en recherche désormais appréciée selon des indicateurs bibliométriques et ils crouleraient sous les demandes de participation à l'appréciation de l'activité de collègues, lors de comité de promotion ou d'octroi de financements de recherche, par exemple. L'évaluation des acquis des étudiants n'échappe pas à cette suspicion généralisée : à la suite de la massification des effectifs étudiants, elle est devenue une composante de plus en plus lourde du métier d'enseignant, accaparant jusqu'à 14 % du temps de travail (Romainville, 2002a). Bref, le malaise est palpable et l'impression générale est celle d'une *sur-évaluation*, d'un monde de l'enseignement supérieur envahi, étouffé, voire perverti par une évaluation omniprésente et débridée.

Sans que ce soit toujours explicite, cette invasion de l'évaluation est souvent présentée comme une nouveauté, un phénomène récent dont l'ampleur serait particulièrement inquiétante dans l'enseignement supérieur. L'extrait suivant (tiré de Zarka, 2009, p. 113) en constitue une excellente illustration :

> Il semble que, malgré les mises en garde de ces dernières années[102] venant de différents côtés, l'installation de dispositifs d'évaluation s'opère actuellement

100. Entendue en termes de productivité et de rentabilité, dans une logique comptable de plus en plus prégnante en matière d'appréciation des actions publiques.
101. Cf. le numéro 37 de la revue CITÉS consacrée, en 2009, à « L'idéologie de l'évaluation. La grande imposture ».
102. C'est nous qui soulignons.

dans tous les secteurs de la société et les institutions : l'hôpital et le système de santé, les institutions d'éducation et de formation en général, les universités et la recherche en particulier [...].

Or, un regard historique rapide montre aisément que l'évaluation s'est de tout temps pratiquée dans l'enseignement supérieur et sous des formes multiples. On pourrait même prétendre qu'évaluer est constitutif de la profession. En atteste d'abord, en ce qui concerne l'évaluation des acquis des étudiants, le rituel des examens pour l'obtention du diplôme, rituel présent dés la création des universités au XII[e] siècle. Les premières universités ont en effet mis en place rapidement un système complet d'examens et de grades (baccalauréat, licence, maîtrise ou doctorat), notamment pour l'octroi du tout premier diplôme universitaire, la *licencia docendi* (Charles & Verger, 1994 ; Renaut ; 1995). Un des enjeux majeurs, pour les universités, de l'instauration de ce système d'examens et de grades était d'ailleurs de s'emparer du pouvoir de validation des compétences intellectuelles des étudiants : elles seraient désormais reconnues par des maîtres sur la base de la maîtrise d'un certain nombre de connaissances et de compétences (dont la fameuse capacité de *disputatio*) et non plus par des chanceliers externes à l'établissement sur une base arbitraire. On retrouve ici le lien très fort qui unit l'évaluation au pouvoir : la formalisation interne d'une pratique d'évaluation permet à ceux qui la mettent en œuvre de s'émanciper d'un pouvoir hiérarchique externe et donc de l'arbitraire, mais cette pratique d'évaluation leur confère alors un nouveau pouvoir. D'ailleurs, les droits élevés acquittés par les étudiants pour pouvoir présenter les examens constituaient une des principales ressources financières des universités naissantes. On retrouve actuellement, dans d'autres domaines et notamment celui de l'accréditation, ce trio infernal qui unit, pour le meilleur mais le plus souvent pour le pire, l'évaluation, le pouvoir et l'argent...

Pour la petite histoire, cette position de pouvoir a ensuite fait l'objet de critiques acerbes. En particulier, la validité et plus prosaïquement le sérieux des examens universitaires ont constitué un sujet fréquent de railleries, comme dans le passage suivant tiré des *Mémoires de ma vie* de Charles Perrault, publié en 1759 (Thélot, 2001, p. 302) :

> Un valet qui vint nous parler à la fenêtre, ayant su ce que nous souhaitions [une licence] nous demanda si notre argent était prêt. Sur quoi, ayant répondu que nous l'avions sur nous, il nous fit entrer et alla réveiller les docteurs, au nombre de trois, qui vinrent nous interroger avec leur bonnet de nuit sous leur bonnet carré. (...) Un de nous, à qui l'on fit une question dont il ne me souvient pas répondit hardiment : *Matrimionium est legitima maris et foeminœ conjuctio, indivuduam vitae consuetudinem continens*, et dit sur

ce sujet une infinité de belles choses qu'il avait apprises par cœur. On lui fit ensuite une autre question sur laquelle il ne répondit rien qui vaille. Les deux autres furent ensuite interrogés et ne firent pas beaucoup mieux que le premier. Cependant, ces trois docteurs nous dirent qu'il y avait plus de deux ans qu'ils n'en avaient interrogés de si habiles et qui en sussent autant que nous. Je crois que le son de notre argent, que l'on comptait derrière nous pendant que l'on nous interrogeait, servit de quelque chose à leur faire trouver nos réponses meilleures qu'elles n'étaient.

S'agissant de l'évaluation de la recherche et de l'appréciation par les pairs des publications, des formes d'évaluation en aveugle sont avérées dès le début du XIII^e siècle même si elles ne se généraliseront qu'après la seconde guerre mondiale avec le développement de la *big science*[103]. À titre illustratif, l'extrait suivant d'un éditorial d'une des plus anciennes revues scientifiques (*Medical Essays and Observations*) publiée par la *Royal Society of Edinburgh* prône, dès 1731, le recours à l'expertise anonyme par les pairs (cité par Milard, 2010, p. 26) :

> Les mémoires envoyés par correspondance sont distribués d'après leur thématique aux membres de la Royal Society qui sont les plus compétents en la matière. Leur identité n'est pas connue de l'auteur.

Bien sûr, le terme même d'*évaluation*, dans son sens actuel, est plutôt récent, même si son étymologie remonte au XIV^e siècle où il apparaît, dans le domaine monétaire, pour désigner l'action d'évaluer, l'*évaluacion*. C'est ainsi que l'on ne parle que depuis peu d'évaluation des acquis des étudiants, même si la pratique correspondante date, en réalité, comme indiqué ci-dessus, de la création des premiers établissements d'enseignement supérieur. Simplement, on parlait auparavant d'examen, d'épreuve ou ultérieurement de contrôle des connaissances. En quelque sorte, les enseignants du supérieur pratiquaient l'évaluation comme Monsieur Jourdain faisait de la prose, c'est-à-dire sans toujours étiqueter leurs nombreuses et diverses pratiques de jugement sous le terme d'évaluation. Mais, si les pratiques d'évaluation sont anciennes et consubstantielles au métier d'enseignant et de chercheur dans l'enseignement supérieur, comment expliquer alors qu'elles fassent actuellement l'objet de tant de critiques et de méfiance ?

103. Ce terme est utilisé par les historiens anglophones des sciences pour désigner l'évolution considérable que les disciplines scientifiques ont connue dans les pays industrialisés à l'issue de la seconde guerre mondiale. D'une science se développant à coups de contributions spontanées, individuelles et très localisées, on serait passé à une science se fondant sur de larges projets d'équipes soutenus par des agences de recherche nationales ou internationales.

2. LES RACINES DU DIVORCE

L'explication de la brouille qui semble présider aux relations entre évaluation et enseignement supérieur est à chercher du côté des modalités selon lesquelles se pratique désormais l'évaluation, ces modalités ayant considérablement évolué depuis deux décennies. Autrement dit, les critiques ne portent pas tant sur la *fonction* d'évaluation elle-même – fonction à laquelle les acteurs de l'enseignement supérieur sont habitués et dont ils reconnaissent d'ailleurs le bien-fondé – mais plutôt sur les *formes* que l'évaluation prend désormais et qui leur semblent à la fois peu pertinentes, inutilement lourdes, grandement consommatrices de ressources et porteuses de sérieuses dérives diverses.

2.1 Une fonction peu remise en cause

En appui de notre hypothèse, observons que même les plus virulents des dénonciateurs de l'actuelle *sur-évaluation* s'empressent (presque) toujours de signaler que leurs critiques ne s'adressent ni au bien-fondé ni aux fonctions sociales du jugement évaluatif, mais qu'elles visent plutôt les modalités selon lesquelles il est actuellement produit. Ainsi, un certain accord sur la nécessité théorique de la présence de nombreuses formes d'évaluation dans le fonctionnement de l'enseignement supérieur peut être relevé. Commençons par le plus consensuel. Personne ne va bien sûr remettre en cause la fonction de certification liée à l'évaluation des acquis des étudiants : les enseignants estiment qu'il fait partie intégrante de leur métier d'apprécier le degré d'acquisition des connaissances et des compétences de leurs étudiants ; ils le revendiquent même. Aux yeux de tous, il s'agit à l'évidence d'une des fonctions essentielles de toute formation, *a fortiori* de l'enseignement supérieur qui se trouve en bout de course du système éducatif et qui certifie l'acquisition par les étudiants de ressources telles qu'ils seront en mesure d'exercer au mieux et au bénéfice de la société une ou plusieurs professions.

Par contre, les critiques pleuvent sur les récentes évolutions des pratiques d'évaluation des acquis : alourdissement considérable des examens à la suite de la massification et appauvrissement de la qualité des épreuves qui en découle (e.g. le recours aux Questionnaire à Choix Multiples) ; injonction faite aux enseignants de clarifier à l'excès leur « contrat didactique » et de justifier leurs notes et leurs décisions ; intervention d'acteurs externes dans le processus (par exemple, les enquêtes nationales ou internationales de type PISA) ; injonction à adapter les épreuves aux nouvelles finalités de formation (e.g. l'approche par compétences) ou aux nouvelles conditions d'exercice du métier (e.g. le développement d'épreuves intégrées collectives).

Les acteurs de l'enseignement supérieur se contentent-ils prudemment de s'accorder sur la nécessité de l'évaluation, à la condition qu'ils

n'en soient pas l'objet ? Pas du tout. Ainsi, l'évaluation des productions scientifiques par les pairs est très largement acceptée par la communauté des chercheurs et chacun se plie, bon gré mal gré, aux évaluations, parfois cinglantes pourtant, d'arbitres anonymes à propos de ses propositions d'articles ou d'ouvrages. On y voit même la condition *sine qua non* de l'authentique débat scientifique et l'assurance d'un mécanisme interne, rigoureux et collectif de contrôle et donc de garantie de validité des connaissances ainsi produites.

Encore une fois, ce qui va susciter les critiques est davantage de l'ordre des formes particulières que cette évaluation de la recherche a eu tendance à revêtir ces dernières décennies. Certains regretteront ainsi la réduction de cette évaluation à quelques indicateurs quantitatifs bibliométriques jugés en définitive peu représentatifs de la réelle qualité de la recherche. D'autres critiqueront le recours à des standards internationaux en la matière qui seraient, dans certains domaines de la recherche en sciences humaines et sociales, incompatibles avec une approche culturellement située de l'investigation scientifique, au risque de connaître une harmonisation réductrice des processus et surtout des paradigmes de recherche (Beauvois, 2009).

Même si l'unanimité est ici moins évidente, le principe de l'évaluation institutionnelle des établissements d'enseignement supérieur et plus généralement des politiques en la matière est lui-même peu contesté. En ouvrant en tant que président de l'Université Paris-Descartes le colloque dont le présent ouvrage est le prolongement, Axel Khan rappelait que le principe même de l'évaluation constitue la principale alternative à l'arbitraire et qu'il est au cœur du processus démocratique : c'est parce que l'on prend la peine de mesurer l'atteinte des propositions politiques faites par les uns et par les autres que la discussion démocratique peut avoir lieu sur des bases sereines et connues de tous. Dans le cas contraire, on est exposé à l'arbitraire ou au « pouvoir des chefs ». S'agissant de l'évaluation des établissements, Compagnon (2003, p. 2) attire l'attention, dans le même sens, sur le fait que la différenciation entre établissements était de toute manière à l'œuvre dans les mentalités collectives et qu'elle pouvait se révéler sauvage et approximative en l'absence d'évaluation systématique :

> Sans évaluation institutionnelle explicite, publique, aisément accessible à tous les acteurs, la différenciation tend à être assurée par des procédures parallèles, médiatiques par exemple, comme les palmarès des universités fournis par les magazines, conduisant les étudiants et leur famille à une attitude consumériste, c'est-à-dire à faire leur marche de formation sur la base de critères approximatifs.

Au final, on trouvera donc peu d'opposition au principe même de l'évaluation, en tant que démarche de recueil systématique et organisé d'informations permettant de poser un jugement de valeur sur une réalité à

apprécier, qu'il s'agisse de la compétence d'un étudiant, de la recherche d'un collègue ou du fonctionnement d'un département et d'un établissement. Là où le bât blesse, c'est dans la manière de mettre en place cette démarche et en particulier dans les changements importants des formes et modalités que les différents types d'évaluation ont connus ces dernières années.

2.2 Des transformations de formes importantes et contestées

Si l'on souhaite mieux identifier les sources de la méfiance dont l'évaluation dans l'enseignement supérieur fait actuellement l'objet, il nous faut donc préciser quelles sont les transformations qu'elle a subies et en quoi et surtout pourquoi ces transformations ont été la source de réticences et de contestations.

Les procédures d'évaluation sont d'abord devenues beaucoup plus **formalisées, normalisées et standardisées**, alors qu'elles avaient longtemps revêtu un caractère largement informel. L'évolution est nette dans tous les domaines. Ainsi, les procédures d'évaluation des acquis des étudiants ont été codifiées, comme le montrent, par exemple, les chartes de l'évaluation, désormais disponibles sur les sites de nombreux établissements. Des règles communes ont été édictées sur la part minimale du contrôle continu, sur l'information faite aux étudiants quant aux critères d'évaluation, sur les significations et les proportions attendues des différentes mentions... Dans certains pays, des réglementations assez strictes en matière de contrôle des connaissances ont été prises par décret ou via des lois, comme les règles de la fameuse et très discutée « compensation » en France ou les normes de réussite partielle d'une année en Belgique francophone. Ce qui était réglé auparavant selon des coutumes et des habitudes pour l'essentiel informelles est désormais régi par un arsenal formalisé et standardisé, parfois assez lourd et très contraignant.

Il en va de même pour l'évaluation de la recherche. Elle se fondait auparavant sur des jugements informels et se réalisait « par et entre pairs ». Les procédures en la matière se sont également formalisées et codifiées ; le recours à des techniques standardisées (notamment de quantification bibliométrique) est de plus en plus fréquent. Ainsi les revues publient désormais systématiquement leurs critères d'évaluation et il n'est plus demandé aux arbitres de rendre un jugement global et informel (du type « à accepter avec ou sans modification » ou « à refuser »), fût-il justifié, mais ils sont invités à cocher des cases ou à remplir une fiche d'avis sur des échelles d'appréciation, le comité de rédaction de la revue procédant ensuite à un comptage des points obtenus par la proposition.

Il en est de même pour les procédures d'acceptation des communications à des colloques, des logiciels spécialisés étant d'ailleurs apparus sur

le marché (e.g. *Conftool*). S'agissant de l'appréciation des chercheurs, les procédures se sont encore davantage technicisées puisque il est désormais possible de se faire une (bonne ?) idée de la qualité de tel ou tel « publiant » à partir d'un certain nombre d'indices bibliométriques, dont le célèbre indice *h* censé quantifier sa « productivité » scientifique en fonction du niveau de citation de ses publications. Bien évidemment, le risque est de perdre en validité et en pertinence de la mesure ce que l'on aura gagné en automaticité et simplicité du calcul, nous y reviendrons.

Une deuxième évolution importante est que de **multiples acteurs extérieurs se sont immiscés dans les procédures d'évaluation** qui jusque-là étaient principalement restées aux mains des enseignants-chercheurs eux-mêmes. Cette évolution est moins perceptible en ce qui concerne l'évaluation des acquis des étudiants, même si de plus en plus de règles externes s'imposent à tous et si des épreuves standardisées et externes commencent à voir le jour, y compris au niveau international[104].

Cette évolution est par contre radicale dans le domaine de l'évaluation de la recherche et encore davantage dans l'évaluation institutionnelle. Alors que l'évaluation de la recherche était exclusivement réalisée « par et entre » des chercheurs, elle est de plus en plus fréquemment organisée par des instances qui échappent en partie aux acteurs[105], au point qu'elle est parfois mise en œuvre par des entreprises privées, extérieures aux établissements d'enseignement supérieur, telles que Thomas Reuter, Elsevier et Science-Metrix pour la production de données bibliométriques.

C'est encore pire, si l'on peut dire, pour les évaluations institutionnelles : la plupart du temps, elles sont commanditées par des acteurs externes et notamment par les pouvoirs publics qui financent les établissements. Ceux-ci insistent par ailleurs pour que l'évaluation prenne en compte des aspects qui étaient considérés jusque-là comme exogènes au fonctionnement même des établissements d'enseignements supérieur, comme les demandes des milieux économiques et sociaux, l'insertion ultérieure des étudiants, la pertinence sociale de la recherche... À propos de ces questions, les autorités publiques estiment que des acteurs extérieurs au « petit monde » du supérieur doivent désormais être partie prenante du processus même d'évaluation, en y jouant un rôle explicite.

La troisième évolution a trait au **caractère de plus en plus public de l'évaluation et de ses résultats.** Alors qu'elle restait bien souvent

104. Après PISA, un nouveau programme de l'OCDE (AHELO) se propose de mesurer, dès 2013, les performances des étudiants et des universités, dans le but de fournir des « données sur la pertinence et la qualité de l'enseignement et de l'apprentissage dans le supérieur ».

105. Ou en tout cas à une grande partie d'entre eux s'ils ne jouent pas le rôle d'experts auprès de ces instances.

confinée à la sphère privée, l'évaluation a acquis, à la suite d'ailleurs de sa formalisation et de l'immixtion d'acteurs extérieurs, un caractère public plus marqué. Des « palmarès » des meilleurs établissements surgissent dans la presse spécialisée ou non, les indices bibliométriques sont accessibles à tous, les classements de revue aussi. Cette évolution est particulièrement observable dans le domaine de l'évaluation des enseignements et des programmes. On a de tout temps évalué les enseignements et les programmes, mais de manière informelle et implicite et sans que les résultats fassent l'objet de diffusion large. Tout change lorsque les résultats aux évaluations des enseignements par les étudiants remontent aux autorités de l'établissement ou sont utilisés pour identifier les forces et les faiblesses d'un programme. Il s'agit d'ailleurs d'une des principales sources de résistance des enseignants face à l'évaluation de leurs enseignements par les étudiants : la démarche suppose en effet que l'acte d'enseignement ne soit plus considéré comme relevant du domaine privé de l'enseignant, à gérer par lui seul dans la pénombre d'un amphithéâtre et dans l'intimité de sa relation avec les étudiants, mais soit conçu comme une contribution à une action collective de formation, organisée autour d'objectifs et de principes communs, discutés et élaborés collégialement (Romainville & Coggi, 2009). Il en va de même lorsque les résultats de l'évaluation institutionnelle sont rendus publics par les agences d'évaluation, comme l'AERES qui publie en accès libre sur son site les rapports d'évaluation des établissements, des unités de recherche, des écoles doctorales, des formations et des diplômes.

Enfin, la dernière évolution, et non des moindres, concerne les effets de l'évaluation et en particulier **ses incidences de plus en plus lourdes sur les personnes, les départements et les établissements**. Tant que l'on évaluait la qualité d'un article pour décider de sa publication ou d'un enseignement pour espérer susciter son amélioration, il n'y avait pas péril en la demeure. Mais dès lors que les résultats de l'évaluation interviennent dans des décisions relatives aux carrières des personnes ou au financement des départements et des établissements, les enjeux sont d'une autre nature et la crispation ne manque pas de s'emparer des acteurs, proportionnellement d'ailleurs au caractère stratégique des décisions auxquelles est désormais liée l'évaluation. Dans une perspective de gestion économe et sélective des moyens alloués à l'enseignement supérieur en fonction de l'efficacité de ses acteurs, l'évaluation se retrouve alors au cœur de processus de différenciation aux implications financières majeures. C'est ainsi que seuls les chercheurs les plus « publiants » pourront encore accéder aux crédits de recherche et l'octroi de financements sera conditionné à l'accréditation des programmes. On comprend que, dans un pareil contexte, la validité des procédures d'évaluation est analysée à la loupe, notamment par ceux qui craignent de sortir, injustement à leurs yeux, perdants de l'opération...

Certes ces transformations radicales des formes prises par l'évaluation dans l'enseignement supérieur sont potentiellement porteuses de dérives majeures, dont l'identification constitue d'ailleurs un des objectifs de la recherche dans le domaine. Mais il faut aussi bien voir que tout n'est pas négatif dans ces évolutions et qu'elles ont notamment contribué à **passer d'une évaluation implicite ou spontanée, voire sauvage à une évaluation explicite et instituée**. Selon le modèle de Barbier (1985) appliqué ensuite par Dejean en 2002 à l'évaluation des enseignements par les étudiants, trois degrés d'objectivation peuvent être distingués dans toute opération d'évaluation.

La première forme d'objectivation correspond à *l'évaluation implicite* dans laquelle le jugement de valeur, dont la production est inhérente à l'acte d'évaluer (nous y reviendrons ci-dessous), ne s'explicite qu'à travers ses effets. Par exemple, des étudiants désertent en masse tel enseignement qu'ils estiment inefficace ; seul est observable l'effet de leur jugement, mais ce dernier n'a pas été explicité ni ses motivations d'ailleurs. L'évaluation du personnel a longtemps fonctionné de la sorte dans l'enseignement supérieur. Après avoir confié quelques premières tâches de gestion (de département ou de laboratoire) à tel collègue et en vertu du fait qu'il ne les avait pas correctement remplies, plus personne n'envisageait de voter pour lui pour d'autres postes à responsabilités. C'était d'ailleurs une excellente manière de ne pas avoir trop de tâches de services à assumer : elles s'amenuisaient d'office si vous aviez négligé les premières que l'on vous avait confiées... Les classements des établissements qui prévalaient de manière informelle dans les mentalités collectives au sein d'un pays ou d'une région procédaient, pour l'essentiel, aussi de l'évaluation implicite. L'attractivité de certains d'entre eux pouvait se mesurer aux demandes d'inscription notamment, mais sans que l'on sache exactement pourquoi et au nom de quoi ils étaient jugés « meilleurs » et surtout sans savoir si ces jugements possédaient une quelconque validité.

Dans la deuxième forme d'évaluation (appelée *spontanée*), le jugement de valeur ne s'explicite qu'à travers sa formulation. Seul est manifeste le jugement auquel aboutit l'évaluateur, mais on ne sait rien de ses fondements, c'est-à-dire des critères et des valeurs au nom desquels il a été produit. L'évaluation des acquis des étudiants s'est longtemps cantonnée à une évaluation spontanée : le processus d'évaluation ne se manifestait en effet qu'à travers son résultat explicite, ramassé bien souvent dans une note sur une échelle de 0 à 20. L'évaluation de la recherche ressortissait parfois de l'évaluation spontanée : l'arbitre était tenu de faire connaître sa décision finale uniquement en termes d'acceptation ou de refus de la proposition d'article. Au début de l'installation des agences qualité, on n'était pas toujours très loin de l'évaluation spontanée, notamment lorsque des « experts », sur la base d'une rapide visite du site et d'une brève rencontre

avec quelques personnes censées représenter l'établissement, fournissaient un certain nombre de jugements sur le rapport d'auto-évaluation préalablement élaboré par l'établissement.

Dans la troisième forme, *l'évaluation instituée*, l'ensemble des éléments du processus d'évaluation (ses effets, les jugements de valeur produits et leurs fondements) passent de l'implicite à l'explicite : le jugement de valeur s'explicite comme le résultat d'un processus social spécifique dont les fonctions, les étapes, les critères et les valeurs de référence sont décrits et sont dès lors susceptibles d'observation, de discussion, voire de contestation.

Par rapport aux formes implicite et spontanée, l'évaluation instituée a l'avantage de la transparence : dès lors que l'ensemble du processus évaluatif fait l'objet d'une explicitation, un des bénéfices de cette forme d'évaluation réside dans la possibilité de mener un débat contradictoire à propos des modalités et des procédures selon lesquelles les jugements sont produits ainsi qu'à propos des critères et donc des valeurs sous-jacentes à ces jugements[106]. Instituer l'évaluation permet notamment d'échapper aux formes sauvages d'évaluation, que ce soit la rumeur, les « on-dit » ou le jugement d'autorité. L'évaluation instituée ouvre donc au débat démocratique, chacun pouvant accéder à l'ensemble du processus d'évaluation et chacun étant en mesure, sur cette base, de critiquer et de remettre en cause à la fois les fondements de l'évaluation, ses procédures et ses résultats. C'est finalement à cette condition que l'évaluation constitue un rempart contre l'arbitraire et une protection à l'encontre du « pouvoir du prince ».

Dans le domaine de l'évaluation de l'enseignement par les étudiants par exemple, ces derniers ont de tout temps jugé leurs enseignants, mais souvent de manière implicite ou spontanée, sur la base de critères non explicités et qui pouvaient finalement être assez étrangers à la qualité de l'enseignement (comme le charisme de l'enseignant). Lorsque l'évaluation des enseignements par les étudiants prend une forme organisée et instituée selon un processus social décidé collectivement, elle offre la possibilité aux enseignants de discuter des critères sur lesquels elle se fonde et leur permet de contester le fait que ces critères rendent compte de manière significative de l'essence de leurs enseignements – que ce soit en termes d'objectifs ou de méthodes – et d'en proposer d'autres jugés plus représentatifs (Romainville, 2010).

106. Une des difficultés – qui rend d'ailleurs assez optimiste ce propos sur la capacité de l'évaluation instituée à constituer un rempart contre l'arbitraire – réside dans le fait qu'il est en réalité souvent illusoire de penser que l'ensemble des éléments d'un processus complexe d'évaluation puissent faire l'objet d'une explicitation complète et définitive. L'exigence d'explicitation doit donc rester réaliste et s'en tenir à porter sur les éléments les plus fondamentaux, sans prétendre que toutes les composantes du processus seront entièrement formulées sans ambiguïté ni équivoque, au risque cependant de voir l'arbitraire resurgir de ce qui serait resté flou...

Comment comprendre alors les résistances auxquelles se heurtent les tentatives d'instauration d'évaluations instituées, que ce soit en matière d'évaluation des enseignements, de programmes ou d'établissements ? Si le passage d'une évaluation implicite ou spontanée à une évaluation instituée, malgré le fait qu'il favorise une plus grande transparence et donc un degré de participation démocratique supérieur, suscite autant de méfiance, c'est sans doute que le flou et l'implicite profitaient à certains et remplissaient des fonctions latentes dans le fonctionnement de l'enseignement supérieur.

La question est donc de savoir « à qui profitaient le flou et l'implicite ? ». Il nous faut pour cela interroger le rôle que remplissait l'évaluation dans l'enseignement supérieur lorsqu'elle était essentiellement cantonnée à une forme implicite ou spontanée.

2.3 Les fonctions latentes de l'évaluation implicite ou spontanée

Globalement, on peut faire l'hypothèse que le caractère diffus et implicite des procédures d'évaluation participait au maintien et à la reproduction des relations de pouvoir et de dominance au sein de l'enseignement supérieur, que ce soit vis-à-vis des étudiants, des collègues ou des autorités publiques.

S'agissant des **étudiants**, Bourdieu et Passeron (1964) ont bien montré que « le système d'exigences diffuses et implicites » auxquels les étudiants étaient confrontés lors de l'évaluation de leurs acquis participait à la reproduction des inégalités sociales dans l'enseignement supérieur :

> En effet, les étudiants des classes cultivées sont les mieux (ou les moins mal) préparés à s'adapter à un système d'exigences diffuses et implicites puisqu'ils détiennent, implicitement, le moyen d'y satisfaire (Bourdieu & Passeron, 1964, p. 113).

Depuis ces travaux pionniers de Bourdieu et Passeron, de nombreuses études ont confirmé l'importance du décodage par les étudiants de ce que Coulon (1997) appelle les « allants de soi » de leur métier. L'inégale prédisposition des étudiants à réaliser correctement ce décodage en fonction de leur origine sociale explique en grande partie pourquoi la démocratisation patine dans l'enseignement supérieur, la réussite y étant conditionnée au « capital culturel » des étudiants, crucial pour l'identification des normes implicites des critères d'évaluation des acquis.

De nombreuses études empiriques ont montré qu'un « contrat didactique[107] » particulièrement opaque transforme l'évaluation des acquis des

107. Le « contrat didactique » est entendu ici comme l'ensemble des règles implicites qui régissent les échanges didactiques entre l'enseignant et ses étudiants et qui précisent tacitement le rôle et les attitudes à adopter par chacun. Le groupe-classe est alors considéré

étudiants en un « jeu du chat et de la souris », les étudiants les plus favorisés ne devant leur salut qu'à un décryptage « sur le tas » des règles tacites et parfois contradictoires de ce contrat. Par exemple, Williams (2005) a comparé la manière dont des étudiants de première année en sciences interprètent une série de verbes fréquents dans les amorces des questions d'examen (définir, discuter, expliquer, prédire, rendre compte) avec les définitions recueillies auprès de leurs enseignants. Il observe d'abord des différences non négligeables dans les définitions fournies par les différents enseignants interrogés. Ensuite, pour une majorité de ces verbes, moins de la moitié des étudiants évoquent, en fin d'année, des définitions compatibles avec celles de leurs enseignants. Une des explications de ce résultat réside dans le fait que les définitions recueillies auprès des enseignants (surtout les plus âgés) s'éloignent assez fortement des définitions de sens commun, telles qu'elles apparaissent pourtant dans des dictionnaires de bon niveau, sans que ces enseignants n'aient pris la peine d'expliciter le sens précis auquel ils avaient eu recours lors de la rédaction de leurs questions. Le flou et l'implicite ont ici une fonction latente de sélection sociale : l'évaluation des acquis ne se cantonne pas à vérifier en toute transparence si les objectifs d'enseignement ont été atteints par les étudiants, mais elle vise à sélectionner les étudiants les plus aptes de repérer ce qui est tacitement attendu d'eux.

En ce qui concerne les **enseignants**, l'accès et le maintien aux positions de pouvoir résultaient, dans le « tout petit monde » de *l'Homo academicus*, d'un jeu très subtil d'influences, de réputations et de positions dominantes, justifiées ou non sur le plan scientifique (Bourdieu, 1984b). On peut donc ici aussi faire l'hypothèse que le flou et l'implicite autorisaient des positionnements favorables à ceux qui savaient en jouer et l'on comprend alors que ces derniers ne se montrent guère enclins à instituer l'évaluation, de manière, par exemple, à ce qu'elle se fonde sur des critères plus objectifs de productivité scientifique :

> Ce pouvoir [des patrons universitaires] sur les mécanismes de reproduction, et par là sur l'avenir du corps (...) repose sur le contrôle, par la cooptation, de l'accès au corps universitaire, sur les relations de protection et de dépendance durable entre le patron et ses clients, et enfin sur la maîtrise des positions institutionnelles de pouvoir, jury de concours de recrutement, Comité consultatif, conseils de facultés, voire commissions de réforme (Bourdieu, 1984b, p. 139).

Ainsi lorsqu'elle est implicite ou spontanée, une évaluation de demandes de crédits de recherche ou de propositions d'articles scientifiques repose sur la réputation du requérant et sur ses relations, voire sur ses liens

comme une société coutumière, la réussite passant par la maîtrise des règles tacites en usage dans cette société.

de soumission et de subordination avec le « mandarin » autour duquel s'organise l'évaluation. On comprend ici aussi comment l'instauration d'une évaluation instituée et l'imposition de règles communes et connues de tous (évaluation en aveugle, critères affichés, anonymat des arbitres...) risquent de contrecarrer grandement le fonctionnement tacite antérieur et donc de remettre en cause les hiérarchies du passé[108].

Enfin, s'agissant des **établissements** et de leurs relations avec les autorités publiques, l'absence de contrôle externe de leur efficacité a été revendiquée dès la création de l'université moderne, sous la célèbre bannière de la « liberté académique » (Renaut, 1995). À la fondation de l'Université de Berlin (modèle de l'université moderne), von Humboldt revendique que les universités jouissent de cette « liberté académique » en vertu du fait que sont rigoureusement imprévisibles les domaines dans lesquels et les voies par lesquelles se produiront les avancées les plus significatives et les plus porteuses du savoir. Les autorités publiques n'ont alors pas à intervenir dans la gestion de la recherche des établissements d'enseignement supérieur, sous peine de nuire au développement libre des sciences, le seul efficace aux yeux de von Humboldt :

> « On ne peut pas se proposer le savoir comme premier objet sans vouloir aussi et en même temps la vie des idées et le mouvement scientifique le plus libre » : ainsi la liberté académique, qui est à la fois liberté à l'égard de l'État et liberté vis-à-vis de la société, est-elle constitutive de l'Université, puisque l'Université trouve sa spécificité dans sa capacité à être un « établissement effectivement scientifique » et que « la science cesse d'être elle-même dès qu'elle n'est plus cultivée pour elle-même ». (Renaut, 1995, p. 125)

Cet argument a longtemps permis aux universités d'échapper au contrôle tatillon de leurs activités de recherche, même si bien sûr des formes d'évaluations implicites ou spontanées avaient cours, comme en témoigne la concentration *de facto* de crédits de recherche sur tel ou tel établissement ou sur tel ou tel laboratoire. Ici aussi, on mesure facilement combien l'instauration d'une évaluation instituée remettrait en cause les positions de pouvoir acquises et, de manière plus globale, les valeurs d'autonomie, d'indépendance et de liberté auxquelles sont fort attachés les acteurs de l'enseignement supérieur.

108. Ici aussi, on pourrait trouver fort optimiste le fait de penser que l'évaluation instituée puisse, à elle seule, faire disparaître les jeux de pouvoir et de relations privilégiées qui régissent le « petit monde » de l'enseignement supérieur. Néanmoins, elle pourrait sans doute contribuer à restreindre leur influence et à garantir une plus grande explicitation des fondements des décisions prises, sans encore une fois croire au grand soir de la transparence totale, le pouvoir et l'influence pouvant s'exercer selon d'autres formes et notamment via l'imposition de critères et d'indicateurs...

2.4 Les arguments en faveur d'une évaluation explicite et instituée

Par leur caractère largement informel et peu transparent, les pratiques d'évaluation qui avaient cours dans l'enseignement supérieur étaient génératrices de prérogatives importantes pour les privilégiés sachant exploiter le flou : les étudiants issus des milieux favorisés qui se jouaient des pièges de l'évaluation implicite et spontanée des acquis, les « mandarins » qui devaient leur position de pouvoir à des jeux d'influence et de relations assez opaques et les établissements eux-mêmes qui garantissaient leur autonomie de fonctionnement en arguant du fait que toute tentative d'immixtion des autorités publiques dans la gestion de leur recherche serait largement contre-productive.

Ce système apparemment en équilibre a été bouleversé sous la pression de nombreuses évolutions que l'enseignement supérieur a connues dans les quarante dernières années du XX^{e} siècle, période à partir de laquelle l'évaluation et l'enseignement supérieur n'ont plus jamais fait véritablement bon ménage.

La massification de la population étudiante et la constitution en « problème » de l'échec socialement injuste ont rendu nécessaire une réflexion en profondeur sur les pratiques d'évaluation des acquis et en particulier sur ces fameux « systèmes d'exigences diffuses et implicites », en partie responsables de l'absence d'ouverture démocratique de l'enseignement supérieur. L'augmentation considérable de la population étudiante a par ailleurs contribué à soulever de nouveaux problèmes de fidélité et de validité des évaluations et a encouragé le développement de procédures technologiques innovantes telles que les Questionnaires à Choix Multiples corrigés par lecture optique, l'évaluation en ligne ou les plateformes de testing. De plus, le développement de l'approche par compétences a requis l'élaboration de nouveaux dispositifs « alignés » d'évaluation qui lui soient compatibles, les examens traditionnels, ponctuels et restitutifs n'étant pas en mesure de rendre compte de la complexité du développement de compétence : évaluation continue et formative, portfolio, évaluation en situation authentique ou semi-authentique....

Par ailleurs, la mondialisation de l'enseignement supérieur a placé les établissements (les plus prestigieux du moins) dans une situation nouvelle de concurrence au sein d'un grand marché où leur position est objectivée à l'aide d'indicateurs quantifiables, plus ou moins fiables. Ces indicateurs sont censés permettre des comparaisons entre établissements et donc l'élaboration de classements et de palmarès. Il en va parfois de même pour les départements et laboratoires, eux aussi classés sur le plan national.

Enfin, la crise des finances publiques, l'exigence adressée aux établissements de « rendre des comptes » de l'usage de leur financement et la tendance, dans certains pays du moins, à lier ce financement à l'efficacité

de l'établissement ou des départements et laboratoires ont abouti à la création de toute une série d'évaluations institutionnelles supplémentaires et standardisées : évaluation de la qualité ; évaluation des filières de formation, des enseignements et des programmes ; évaluation des laboratoires... Cette inflation des procédures d'évaluation qui relèvent, aux yeux de la plupart des acteurs, plus de la fonction de contrôle que d'un réel souci de promotion de la qualité a été très sévèrement critiquée par ces derniers notamment parce qu'ils redoutaient d'y perdre leur autonomie, ces procédures étant de plus en plus décidées et gérées de l'extérieur, comme indiqué ci-dessus. C'est dans ce contexte et pour ces raisons que la question de la légitimité des acteurs de l'évaluation est apparue sur le devant de la scène.

Après avoir dressé un panorama du nouveau paysage au sein duquel les pratiques d'évaluation prennent désormais place dans l'enseignement supérieur, l'objectif de la partie suivante de ce chapitre est de proposer un inventaire synthétique – sur la base des chapitres précédents mais aussi des conférences et des communications présentées au 23e colloque de l'ADMEE – des questions vives qui se posent dans chaque domaine où est pratiquée une forme d'évaluation et des pistes qui sont actuellement explorées pour produire du savoir sur ces questions et pour tenter de les résoudre.

3. LES QUESTIONS VIVES

3.1 Un modèle général de l'évaluation

La présentation et la discussion des questions qui sont au cœur de la recherche actuelle sur l'évaluation dans l'enseignement supérieur ne peuvent se réaliser en l'absence d'un modèle simple, mais opérationnel de l'évaluation. Sans entrer dans les nombreuses polémiques qui ont entouré cette notion et sur la base de quelques modélisations du processus évaluatif (Barbier 1985 ; De Ketele, 1992 ; Hadji, 1997 ; Madaus, Scriven & Stufflebeam, 1986), on peut considérer qu'évaluer consiste à **apprécier la valeur d'une chose (une action, ses résultats, une personne, un établissement...) en regard d'objectifs, cette appréciation se déroulant en trois étapes.** La première étape[109] de ce processus réside dans le ***recueil systématique, valide et fidèle d'informations appropriées aux objectifs*** dont on souhaite mesurer l'atteinte. C'est la phase d'observation et/ou de recueil de données. C'est durant cette phase qu'interviennent d'éventuelles mesures, qui ne constituent qu'un élément du

109. C'est par commodité que nous parlons d'étape. Il serait plus adéquat de parler de « composante », car dans la réalité du jugement évaluatif, il n'est pas rare que les trois « phases » que nous dissocions ici co-existent et que des allers et retours soient observés entre chacune d'elles. C'est d'ailleurs ce que montre Marthe Hurteau au chapitre 8 à propos de l'évaluation de programme.

processus évaluatif et auxquelles ce processus ne peut donc se réduire. La deuxième étape consiste à ***interpréter les informations recueillies à l'aide de critères***. C'est la phase d'analyse et d'interprétation des données recueillies. Celle-ci débouche sur la troisième étape dite de ***jugement ou de rétroaction*** durant laquelle le processus aboutit soit à l'établissement de conclusions et/ou à des prises de décision, soit à des actions régulatrices.

Ainsi, lorsqu'un enseignent évalue les acquis de ses étudiants à l'issue d'un enseignement, il procède à un recueil standardisé d'informations (phase 1). Par exemple, il soumet, à une même date et dans les mêmes conditions (temps de réponse standardisé, questions identiques ou équivalentes...) tous les étudiants à un examen écrit durant lequel ces derniers produisent des réponses, censées mobiliser de manière représentative les connaissances et compétences que l'enseignement visait à leur faire acquérir. Mais ces informations, les réponses des étudiants, ne parlent pas d'elles-mêmes. Encore faut-il les analyser à l'aide de critères (étape 2), ces derniers découlant des objectifs de formation. Même analysées de manière rigoureuse en regard de critères, les productions des étudiants n'aboutissent pas non plus mécaniquement à un jugement ou à une décision finale (étape 3 de l'évaluation certificative). Il faudra encore que l'enseignant parcourt un arbre de décisions, explicite ou implicite, lui permettant d'établir un jugement final (e.g. affirmer que la compétence est acquise ou non) à partir de ses analyses. En évaluation formative, cette troisième étape consiste à puiser dans l'analyse des réponses des étudiants des pistes de rétroaction possibles, que ce soit à destination de l'enseignant pour adapter son enseignement ou à destination des étudiants pour affiner et corriger leurs apprentissages (cf. chapitre 3).

Précisons qu'ainsi définie l'évaluation comprend nécessairement une part de subjectivité et que le débat entre objectivité et subjectivité en matière d'évaluation est particulièrement vain et stérile (Romainville, 2012). En effet, évaluer n'équivaut pas à enregistrer de manière externe et objective l'atteinte d'objectifs, comme le thermomètre indique la température, sans l'intervention de l'homme. Au contraire, évaluer revient à construire un point de vue sur des réalités censées représenter l'atteinte d'objectifs. À ce titre, elle comprend nécessairement une part de subjectivité et ce, aux trois étapes évoquées ci-dessus.

Observons également qu'un élément central de l'évaluation ainsi définie consiste à déterminer les personnes qui sont en charge de la définition des objectifs en regard desquels l'évaluation se réalisera. Dans un premier cas de figure, ce sont les acteurs eux-mêmes qui ont défini leurs objectifs. L'évaluation, même si elle est réalisée par un agent externe, consiste alors à apprécier l'atteinte de ces objectifs et notamment la cohérence entre ces derniers et les actions des acteurs. C'est bien souvent sur ce modèle que fonctionnaient les premières agences d'évaluation de

la qualité dans l'enseignement supérieur. Il était demandé aux établissements de réaliser un rapport d'auto-évaluation et le comité d'experts était censé analyser ce rapport et notamment la cohérence entre les objectifs annoncés, les actions entreprises et les résultats obtenus. Mais d'autres cas de figure existent. Il se peut notamment que l'évaluation se réalise selon des critères qui relèvent d'objectifs qui n'étaient pas nécessairement à la base des actions des acteurs. C'est un des reproches fréquemment adressés aux épreuves PISA qui prétendent mesurer l'efficacité des systèmes éducatifs sur la base de mesures de l'atteinte d'objectifs dans lesquels les enseignants ne se reconnaissent pas, tant le type d'épreuves et surtout le type de compétences mesurées (e.g. des compétences pragmatiques liées à la vie quotidienne pour la lecture) sont en définitive assez éloignés de leurs objectifs d'enseignement (Romainville, 2002b). On pourrait adresser le même reproche aux classements internationaux des établissements d'enseignement supérieur. Lorsque le classement de Shanghai fait du nombre de prix Nobel un critère majeur, il sous-entend que la recherche a pour objectif de produire des prix Nobel et que cet objectif serait valable pour l'ensemble des établissements mis en compétition, ce qui bien sûr n'a pas beaucoup de sens. De plus, cet objectif est-il réellement représentatif des finalités de production de savoirs qui constituent le cœur même de cette mission des universités ? On en doute aussi...

3.2 L'évaluation des acquis des étudiants

Parmi l'ensemble des travaux actuels sur l'évaluation des acquis des étudiants, trois tendances majeures peuvent être dégagées. Une première série d'études porte sur les qualités attendues de cette évaluation, en particulier de sa phase de recueil d'informations, lorsqu'elle porte sur des acquis de niveau de complexité élevé et qu'elle prend des formes spécifiques à l'enseignement supérieur. Un deuxième groupe de travaux a trait aux effets des pratiques d'évaluation sur les agents et sur les bénéficiaires du processus évaluatif. Enfin, une troisième série de contributions s'intéresse au développement de pratiques innovantes d'évaluation des acquis (telle que l'évaluation formative et l'évaluation des compétences), aux défis posés par ces pratiques ainsi qu'aux conditions favorables à leur mise en place.

3.2.1 Les qualités de l'évaluation d'acquis spécifiques

Poursuivant des finalités spécifiques souvent à caractère professionnalisant marqué, l'enseignement supérieur rencontre des problèmes docimologiques particuliers, liés notamment à la complexité des acquis sur lesquels porte l'évaluation. Ainsi, les qualités attendues traditionnellement de toute évaluation (validité, fidélité, transparence...) y prennent une coloration particulière.

C'est notamment le cas des nombreux programmes de formation qui se sont emparés du concept de *compétence* et qui prétendent, dans la foulée, évaluer et certifier le développement et l'acquisition de compétences. Or la compétence, objet multidimensionnel s'il est en, ne s'exerce qu'en situation, bien souvent complexe de surcroît. En toute cohérence et comme le montre Linda Allal au chapitre 1 à partir d'exemples convaincants, l'évaluation est alors préférentiellement réalisée en situation authentique ou semi-authentique, ce qui ne va pas sans poser de redoutables problèmes de validité : mesure-t-on bien la part qu'a prise la formation dans le développement de la compétence ? Quel est l'impact des caractéristiques de la situation sur la manifestation de la compétence ?... C'est ainsi que de nombreux travaux cherchent à apprécier la fidélité et la validité des méthodes d'évaluation des compétences par tâches complexes, voire en stage et notamment des méthodes d'interprétation des performances des étudiants censées y refléter la maîtrise de compétences (Coen, 2011 ; Leclerc, Feirreira & Mallet, 2011 ; Monnard & Luisoni, 2011). Bien souvent, ces travaux aboutissent à la conclusion qu'une évaluation portant sur de tels acquis complexes requiert d'être réalisée sur la base de critères et d'indicateurs préétablis et discutés tant par les évaluateurs que par les évalués.

Certains vont jusqu'à se demander si l'enseignement supérieur ne manque pas de modestie et de réalisme lorsqu'il prétend certifier la maîtrise de compétences mises en œuvre au moins partiellement en situation professionnelle. Est-on réellement en mesure de réaliser une telle certification ? Quelle serait la légitimité, ne fût-ce qu'aux yeux des étudiants, de la validation de telles compétences par des enseignants qui bien souvent n'ont pas une grande connaissance des milieux professionnels ? N'appartient-il pas en définitive à ces milieux eux-mêmes de procéder à une telle reconnaissance ? L'enseignement supérieur ne doit-il pas en rester plus prudemment à certifier la maîtrise de ressources nécessaires à la mobilisation de compétences et à évaluer les premières formes de mobilisation, encore réalisées dans un cadre formatif, c'est-à-dire les prémices ou les indices de la mise en œuvre de compétences professionnelles ?

À l'appui de ces réserves, Anne Jorro, Renée Brocal et Nadine Postiaux montrent, au chapitre 4, que le projet d'évaluer la professionnalité est ambitieux et en partie irréaliste et qu'il vaut donc mieux s'en tenir à évaluer ce qu'elles appellent la « professionnalité émergente », c'est-à-dire des traces de professionnalité dont l'étudiant fait déjà preuve dans le cadre de sa formation et de ses stages. Elles montrent aussi combien ce type d'acquis complexes requiert une évaluation sans doute davantage partagée (avec les acteurs du secteur professionnel qui accueillent les étudiants en stage notamment) et des outils spécifiques (comme le portfolio), permettant de rendre compte du développement progressif, idiosyncratique et situé de ce

type de compétences. Mais ces outils ne sont pas non plus exempts d'ambiguïté : les auteures de ce chapitre font état des réticences exprimées par des étudiants lorsqu'ils sont invités à consigner dans leur portfolio les difficultés qu'ils ont rencontrées lors de la mise en œuvre des compétences, sachant que le même outil constituera aussi la base de l'évaluation certificative.

D'autres études s'interrogent sur la validité et la fidélité des jugements auxquels aboutissent des formes d'évaluation assez spécifiques à l'enseignement supérieur et à son fonctionnement, comme les décisions collectives de jury, notamment de VAE (Nkeng & Uebersfled, 2011) et d'attribution de bourses doctorales (Tourmen & Droyer, 2011) ; la décision informelle, mais très chargée symboliquement d'autorisation de soutenance (Vialle, 2011) ou l'appréciation de dossiers d'enseignement déposés par des enseignants à des fins de promotion (Wouters, Frenay & Laloux, 2011).

Enfin, le développement de la Valorisation des Acquis de l'Expérience (VAE) soulève de très nombreuses questions d'évaluation. Au chapitre 2, Marie-Christine Presse montre que, derrière une procédure administrative qui se banalise dans l'enseignement supérieur, la VAE constitue une démarche d'une très grande complexité et qui fait apparaître des défis majeurs en termes d'évaluation. Par exemple, si des compétences se développent à l'évidence par l'expérience, les personnes qui cherchent à les faire reconnaître en regard d'exigences académiques doivent en rendre compte par une analyse réflexive de leurs pratiques professionnelles. Cette mise à distance des acquis de l'expérience suppose notamment de fournir une description décontextualisée des compétences acquises. Or, il s'agit là d'un savoir-faire de haut niveau, qui passe entre autres par la maîtrise d'un genre d'écrit particulier allant parfois bien au-delà de la compétence qui en fait l'objet. Une personne pourtant reconnue comme compétente par ses pairs et sa hiérarchie pourrait ainsi se trouver en grande difficulté de rendre compte par écrit des compétences qu'elle a acquises à travers son expérience professionnelle. Ce qui est évalué n'est alors plus de l'ordre des acquis, mais ressortit plutôt à une capacité de les objectiver dans un écrit spécifique répondant à des normes en partie arbitraires.

De manière plus générale d'ailleurs, la VAE réinterroge les fondements de l'évaluation et de la certification et interpelle jusqu'à la cohérence pédagogique des programmes de formation supérieure. Ainsi, le fonctionnement de la VAE suppose que l'on soit capable d'établir des relations d'équivallence entre des compétences acquises par l'expérience et des parties de programmes académiques. Cette opération n'est possible qu'à la condition que ces programmes aient eux-mêmes identifié et répertorié les compétences qu'ils visent à faire acquérir aux étudiants, ce qui est loin d'être toujours le cas.

La VAE pose également la question de la relation entre les contenus des formations et les savoirs mis en œuvre dans l'activité : les contenus

enseignés participent-ils réellement au développement de l'expertise professionnelle ? Y sont-ils strictement obligatoires ? Y suffisent-ils ? Par ailleurs, l'expertise professionnelle ne peut-elle pas se développer selon d'autres voies, moins formelles et plus expérientielles ? C'est donc toute la pertinence des contenus des programmes de formation en regard de leur capacité à être mobilisés en vue d'une efficacité professionnelle qui est réinterrogée de manière fondamentale par la VAE. Cette dernière peut donc être considérée comme un véritable cheval de Troie, introduisant subrepticement dans l'enseignement supérieur des questions pédagogiques fondamentales qui la dépassent largement.

3.2.2 Les effets et bénéfices de l'évaluation

Une deuxième voie de recherche s'intéresse à l'impact des pratiques d'évaluation des acquis tant sur ses agents (les enseignants) que sur ceux qui en font l'objet (les étudiants). On cherche ainsi à identifier les effets de l'évaluation sur les pratiques étudiantes, sur leur approche d'apprentissage ou sur leurs stratégies d'étude. On se demande aussi si les dispositifs d'évaluation formative incitent effectivement les enseignants à procéder à une régulation de leurs enseignements. On s'interroge enfin sur l'impact global et à long terme de l'évaluation via notamment l'étude du retentissement subjectif qu'ont sur des adultes des épisodes évaluatifs marquants (Baudouin & Vanini, 2011).

On se pose par ailleurs la question de savoir si les bénéficiaires de l'évaluation formative adhèrent à ses valeurs sous-jacentes. Et ce n'est pas toujours le cas : une étude rapporte que les étudiants adhérent en définitive assez peu au paradigme de l'évaluation formative et restent demandeurs d'évaluation discriminante (et de notes !), adoptant en cela une idéologie méritocratique et de sélection, à mille lieux du discours lénifiant de leurs enseignants sur les potentialités formatrices de l'évaluation (Rodrigues, Peralta & Nunes, 2011).

Des effets plus globaux sont aussi analysés. Ainsi, en regard des objectifs d'égalité des chances et d'accroissement de l'accessibilité à l'enseignement supérieur au nom desquels la VAE a été mise en place, des études cherchent à identifier plus précisément quel type d'adultes en reprise d'études bénéficie effectivement des procédures de VAE et à savoir ainsi si ces procédures contribuent à adapter l'enseignement supérieur aux besoins et caractéristiques de populations socialement diversifiées (Maes *et al.*, 2011).

3.2.3 Les pratiques évaluatives innovantes

La troisième voie de recherche a trait au développement de pratiques évaluatives innovantes : leurs fondements, leurs effets, leur efficacité et leurs conditions d'implantation. Pour l'essentiel, deux axes d'innovation

sont explorés : celui, d'une part, de l'évaluation formative qui cherche, comme l'indique Linda Allal au chapitre 1, à jeter des ponts « constructifs » entre trois activités restées largement disjointes dans l'enseignement supérieur (la formation, l'apprentissage et l'évaluation) et celui, d'autre part, de l'évaluation de compétences et d'objectifs de haut niveau taxonomique.

Le mouvement vers l'**évaluation formative** est assez net dans les paliers primaire et secondaire des systèmes éducatifs, mais l'enseignement supérieur en est resté longtemps éloigné. Linda Allal montre au chapitre 1 que des transformations sont cependant en cours en son sein et que des pratiques d'évaluation formative s'y développent actuellement avec une certaine intensité. Le chapitre 3 en fournit un bel exemple : Valérie Wathelet et Sandrine Vieillevoye analysent, à l'aulne des principes de l'évaluation formative, une expérience de détection et de remédiation précoces, auprès d'étudiants de première année, de l'absence de maîtrise des prérequis des formations universitaires. Dans ce dispositif, l'évaluation est au service à la fois de l'apprentissage des étudiants et des pratiques enseignantes. L'étudiant reçoit un feedback formatif sur l'adéquation de ses acquis aux prérequis attendus et bénéficie, au besoin, d'activités de renforcement de son bagage d'entrée. Les enseignants, quant à eux, tirent profit d'une meilleure connaissance du niveau de leur public pour y adapter les premiers enseignements.

S'agissant du pont entre évaluation et apprentissage, l'idée de base de l'évaluation formative est que la rétroaction en cours de route à destination de l'étudiant constitue un facteur clé d'apprentissage. De nombreux auteurs (Carless *et al.*, 2011 ; Nicol & MacFarlane-Dick, 2006 ; Wiliam, 2010) s'accordent sur les caractéristiques d'une rétroaction efficace, telle qu'elle est, par exemple, mise en œuvre dans le dispositif « Passeports » à l'entrée des études universitaires (chapitre 3). La rétroaction doit être précoce et régulière et suivre de près la mesure. Elle doit, par ailleurs, éviter d'être démotivante, ce qui suppose qu'elle soit réalisée sur un ton et transmise dans une forme ni trop critiques ni trop autoritaires. Elle doit concerner la seule performance de l'étudiant et ne pas viser sa personne. Une rétroaction qui oriente de manière efficace l'action de l'étudiant (adaptation de ses stratégies, développement de ses efforts...) doit réunir en outre les qualités suivantes :

- être informative et dès lors ne porter que sur un nombre limité d'aspects des premières performances de l'étudiant ;
- être compréhensible par l'étudiant et donc être exprimée dans des termes qui ne nécessitent pas eux-mêmes que les apprentissages visés aient eu lieu ;
- comprendre des pistes d'amélioration, voire des indications sur les remèdes ;
- être individualisée et ouvrir au dialogue ;

– être suffisamment détaillée (c'est-à-dire expliquer en quoi et pourquoi les performances de l'étudiant répondent ou non aux exigences attendues), sans cependant être trop précise.

Cette dernière qualité est essentielle. En effet, l'évaluation formative vise à long terme à favoriser les processus d'autorégulation de l'étudiant et il convient dans ce sens de ne pas l'habituer à ce que l'enseignant régule à sa place ses processus d'apprentissage. Cette conception de la rétroaction peut être qualifiée de *constructiviste* : le feedback n'est pas un « cadeau » transmis par l'enseignant, mais un message qui doit être interprété par l'étudiant et auquel il doit apprendre à donner du sens. C'est dans cette perspective que Carless et ses collaborateurs (2011) parlent de *feedback durable*, c'est-à-dire centré sur le développement des compétences d'autorégulation de l'étudiant, censées lui apporter à long terme une capacité à tirer lui-même profit des rétroactions futures.

Comme le montre Linda Allal au chapitre 1, les dispositifs d'évaluation formative prennent des formes assez variées dans l'enseignement supérieur. Il peut s'agir de pratiques de contrôle continu ou de la multiplication des étapes et des acteurs de l'évaluation (e.g. présentation orale intermédiaire d'un travail et évaluation par les pairs ; possibilité accordée aux étudiants de procéder à une nouvelle soumission de leur travail). Le portfolio commenté constitue également un outil particulièrement adapté à la fonction formative de l'évaluation : l'étudiant y consigne progressivement des traces pertinentes et probantes faisant foi de l'acquisition de compétences. Il en va de même des grilles d'auto-évaluation, de l'évaluation et la discussion intermédiaire entre pairs et des entretiens d'autoconfrontation avec des étudiants en formation d'enseignants à propos de leurs compétences professionnelles et de leur construction identitaire (Leroy & Beckers, 2011).

Une des difficultés à laquelle ces nouvelles pratiques d'évaluation formative sont confrontées de manière récurrente a trait à leur articulation, nécessaire mais problématique, avec l'évaluation certificative. Dans une approche par compétences en particulier, dès lors que l'évaluation y est continue et recourt à des outils d'enregistrement progressif des indices de maîtrise des compétences, il serait contradictoire d'ignorer, le temps de la certification finale venu, tous les témoignages recueillis en cours de formation, même s'ils l'ont été surtout dans une perspective formative. Mais l'utilisation à des fins de certification de matériaux accumulés à des fins formatives ne va pas sans poser question : quelle part faut-il attribuer aux évaluations formatives intermédiaires dans la décision de certification finale ? Comment s'assurer qu'elles remplissent correctement leur fonction formative si les étudiants connaissent leur rôle ultérieur dans la certification ? Comment, par exemple, assurer l'authenticité de l'auto-évaluation dans ce cadre ? La résolution de ces questions passe par le développement

de dispositifs complexes articulant ces différentes dimensions de l'évaluation (Mottier Lopez & Tessaro, 2011).

La seconde direction empruntée par les pratiques évaluatives innovantes a trait aux défis posés par l'évaluation d'acquis de haut niveau taxonomique et en particulier par **l'évaluation de compétences**. Dans le domaine de l'évaluation des acquis, une des tendances les plus marquantes du 23e Colloque de l'ADMEE réside dans la haute complexité des acquis dont les équipes pédagogiques souhaitent mesurer la maîtrise par les étudiants. Bien au-delà des simples connaissances dont la vérification de l'acquisition a longtemps été la base de l'évaluation dans l'enseignement supérieur, on tente désormais d'appréhender la maîtrise d'acquis nettement plus complexes tels que des savoir-faire (e.g. le calcul de doses en soins infirmiers), des stratégies mentales (e.g. la métacognition des étudiants lors d'un enseignement méthodologique), des postures (e.g. la réflexivité), des jugements (e.g. le jugement moral dans la formation éthique des médecins), des attitudes (e.g. l'identité professionnelle en formation initiale d'enseignants) et bien sûr des compétences, de tous les types de surcroît : compétences académiques disciplinaires, langagières, méthodologiques, transversales, interdisciplinaires et professionnelles, y compris des compétences professionnelles complexes (e.g. les stratégies de gestion de conflits de futurs conseillers pédagogiques).

Dans un contexte d'exigence croissante d'alignement entre les objectifs, les méthodes de formation et l'évaluation, toute la difficulté est de développer des dispositifs évaluatifs qui soient congruents par rapport à ces acquis complexes. S'agissant de l'approche par compétences qui tend à se développer de manière considérable dans l'enseignement supérieur (Romainville, 2007), cette congruence est importante à plus d'un titre. Elle assure d'abord la cohérence pédagogique des formations. Elle envoie par ailleurs un message fort aux étudiants, dont on sait que les pratiques d'études sont véritablement pilotées par les conditions de l'évaluation : c'est bien cette approche qui « comptera » et c'est donc bien sur elle qu'ils ont « intérêt » à aligner leurs stratégies d'apprentissage. Cette congruence revêt une importance accrue dans la formation initiale des enseignants au sein de laquelle le principe d'homologie veut qu'il y ait concordance, pour assurer la crédibilité de la formation notamment, entre la façon dont les étudiants sont évalués et celle par laquelle on préconise qu'ils évaluent leurs élèves.

De nombreuses études cherchent donc à mettre en évidence les critères d'une évaluation des acquis « compétences compatible ». De manière synthétique, ces critères peuvent se rassembler autour de cinq exigences.

Premièrement, on ne peut évaluer que des compétences correctement identifiées et modélisées. Il s'agit du b.a.-ba de toute démarche d'évaluation des acquis : la première opération consiste à délimiter précisément

l'objet à évaluer, ce qui permettra de décider si les buts de l'enseignement ont été atteints, en l'occurrence si les compétences ont été acquises. Et l'exercice n'est guère aisé quand les acquis attendus sont de l'ordre de la compétence. Souvent présentée comme un concept « flou » et multidimensionnel, la compétence peut se définir en première approximation comme une aptitude à mobiliser et à mettre en œuvre de manière efficace un ensemble intégré de ressources (savoirs, savoir-faire et attitudes) permettant, dans une classe de situations, de résoudre un problème ou d'affronter une tâche. La certification de l'acquisition de compétences exige de disposer d'une description des performances qui seront censées rendre compte de cette acquisition ; la rétroaction formative devra, quant à elle, se fonder sur un inventaire des ressources à mobiliser et sur un modèle de fonctionnement de la compétence. Dans ce sens, l'évaluation par compétences suppose des démarches préalables d'élaboration de référentiels précis, mais qui restent maniables et significatifs.

Deuxièmement, une compétence se développe progressivement, selon des schémas d'évolution complexes, faits d'avancées et de retours en arrière. Qui peut par ailleurs prétendre avoir développé à son maximum et à jamais une compétence ? Les compétences s'accommodent donc mal de l'examen certificatif final tel que l'enseignement supérieur l'a longtemps connu : un bref coup de sonde ponctuel, réalisé aléatoirement et arbitrairement tel jour à telle heure et censé dresser le bilan de plusieurs mois d'acquisition. Au contraire, l'évaluation de la maîtrise d'une compétence sera davantage continue et progressive, elle portera sur le long processus de son développement et comprendra des facettes cumulative et formative majeures.

Troisièmement, une compétence ne se manifeste qu'en situation, ce qui requiert que son évaluation se réalise par des mises en situations complexes et ouvertes (c'est-à-dire en regard desquelles aucune solution évidente ne se détache), significatives et authentiques ou semi-authentiques (e.g. les stations d'évaluation formative décrites au chapitre 1). Dans une perspective formative, le seul fait de plonger l'étudiant au sein de situations complexes ne suffira cependant pas à lui faire élaborer des pistes d'amélioration. Encore faudra-t-il observer finement ses actions et analyser (et l'inciter à analyser lui-même) ses gestes, ses hésitations, ses modes d'action et ses raisonnements en regard du modèle de la compétence visée dont le formateur dispose. Ce retour métacognitif sur la mise en œuvre de la compétence est également crucial pour son transfert ultérieur : on sait en effet qu'une compétence ne se transfère à d'autres situations que si elle a fait l'objet d'une décontextualisation minimale : en quoi ce qui a été appris dans telle situation est-il applicable à d'autres contextes plus ou moins proches ? Quels sont les ressources qui ont été nécessaires à la résolution et quelles sont celles qui resteront utiles dans d'autres contextes ? L'agencement de

ces ressources qui s'est révélé efficace dans cette situation sera-t-il valable dans d'autres types de situations ou même dans d'autres situations du même type ?

Quatrièmement, c'est finalement la personne elle-même qui est la mieux placée pour parler de sa compétence. L'évaluation de compétences se tournera donc vers davantage de co-évaluation et vers une évaluation partagée et responsabilisante, accordant au détenteur de la compétence le soin de décrire et de documenter le développement de ses acquisitions.

Enfin, cinquièmement, la mesure de l'acquisition d'une compétence ne se laisse pas aisément réduire à une banale note sur une échelle de 0 à 20, fût-elle critériée. Si l'on pouvait déjà douter sérieusement de la pertinence des notes dans le cadre de l'évaluation de connaissances, leur capacité à rendre compte du développement de compétences est largement sujette à caution. Prétendre qu'obtenir 14 sur 20 signifie que l'on s'est montré deux fois plus compétent que si l'on s'était vu attribuer une note de 7 prête à sourire... Plus fondamentalement, la réduction à une note du processus complexe d'acquisition d'une compétence est outrageusement simplificatrice. L'évaluation de compétences s'oriente donc vers une approche davantage qualitative, descriptive (sur la base d'indicateurs ou d'échelles développementales de la compétence) et surtout informative, comme le montre l'information retournée aux étudiants dans le cadre du projet « Passeports » (chapitre 3). L'essentiel est alors d'accumuler et puis de retourner à l'étudiant des informations sur les ressources non acquises et sur les processus de mobilisation incorrectement ou incomplètement mis en œuvre, ce qui suppose, une fois de plus, de disposer d'un modèle de la compétence. L'évaluation de compétences sera également plus continue que l'examen traditionnel et davantage concertée, c'est-à-dire réalisée en collaboration, d'une part, entre enseignants (e.g. l'épreuve intégrée) et, d'autre part, entre les responsables de la formation et les autres acteurs (e.g. les responsables des stages).

Dans cette perspective d'une évaluation congruente par rapport à l'approche par compétences, de nombreux travaux portent également sur le développement et l'analyse d'outils *ad hoc* : portfolio (notamment électronique), épreuve intégrée, examens en situations authentiques, simulations, études de cas, stations d'évaluation, réseaux de concept, projets...

3.3 L'évaluation des enseignements, des formations et des programmes

Un deuxième champ d'investigation largement exploré actuellement par les études sur l'évaluation dans l'enseignement supérieur a trait à l'évaluation des enseignements, des formations et des programmes. Il s'agit ici de récolter des informations auprès de sources diverses (les enseignants

eux-mêmes, les étudiants, les employeurs...) dans le but d'améliorer les formations et/ou de porter un jugement sur leur qualité, entendue comme leur capacité à atteindre les objectifs qu'elles s'étaient fixés. Compte tenu des résistances auxquelles de pareilles évaluations ne manquent pas de se heurter dans un enseignement supérieur tout acquis à la cause de la « liberté académique », on ne s'étonnera guère que les questions de la validité et de la pertinence des informations récoltées, voire des sources elles-mêmes dans le cas des étudiants, fassent l'objet d'investigations poussées et critiques. On s'interroge également sur les modalités d'interprétation de ces informations et d'élaboration du jugement ainsi sur les fonctions attribuées explicitement ou implicitement à ce type d'évaluation.

3.3.1 *L'évaluation de l'enseignement par les étudiants*

Objet de nombreuses polémiques depuis son apparition dans les années 1960 en Amérique du Nord, l'évaluation de l'enseignement par les étudiants jouit, comme le rappelle le chapitre 6, d'un statut paradoxal et ambigu. Son aspect paradoxal est lié à la contradiction lancinante entre, d'un côté, de nombreuses études assez convergentes qui confirment la validité globale des informations recueillies auprès des étudiants sur certains aspects de leur formation et, d'un autre côté, la méfiance, voire la suspicion que cette forme d'évaluation provoque auprès de nombreux enseignants (Romainville & Coggi, 2009). Sans doute, cette réticence des acteurs trouve-t-elle une partie de son origine dans l'ambiguïté des discours à propos des fonctions de ce mode d'évaluation. Le chapitre 6 rappelle combien, alors que la fonction douce et auto-normée d'amélioration des pratiques pédagogiques est largement mise en avant dans les propos lénifiants qui instaurent et justifient l'évaluation par les étudiants, cette fonction cède *de facto* bien souvent sa place, dans les pratiques effectives, à une autre, plus dure et davantage hétéro-normée, de contrôle. Or, la gestion au quotidien de cette ambiguïté ne va pas sans soulever d'importantes tensions éthiques, notamment en regard des obligations de confidentialité, de transparence et d'accompagnement (Dumont, 2011).

C'est sans doute cet aspect paradoxal et ambigu de l'évaluation de l'enseignement par les étudiants qui explique qu'une partie importante des travaux qui lui sont consacrés explorent la question de sa **fonction** spécifique et de sa **validité**. La question de sa fonction concerne notamment l'articulation – souvent présentée comme une nécessité mais rarement assumée en termes de dispositif – de cette forme d'évaluation avec d'autres types d'évaluation et son intégration au sein d'une démarche qualité globale. Elle touche aussi à l'utilisation des résultats de l'évaluation d'enseignements individuels au sein d'une approche programme collective. La question de sa validité, quant à elle, comprend plusieurs facettes. On s'interroge d'abord sur la légitimité de la source auprès de laquelle les

informations sont recueillies, soit les étudiants : sont-ils en mesure de rapporter de manière valide et fiable des éléments représentatifs de la qualité des enseignements dont ils bénéficient ? Si oui, lesquels et à quelles conditions ? Des biais, tels que l'indulgence des enseignants lors de l'évaluation des acquis, risquent-ils de contaminer à tort leurs jugements ? Si oui, comment les éviter, diminuer leurs effets ou, au moins, en tenir compte lors de l'interprétation des résultats ? De plus, le chapitre 6 montre que la notion de la validité a été étendue et porte désormais aussi sur la capacité du dispositif d'évaluation à produire ses effets attendus et à ne pas engendrer d'effets non-désirés (e.g. l'abattement de jeunes enseignants confrontés à des évaluations trop négatives).

On s'aperçoit alors que, même si elle est souvent présentée comme étant essentiellement au service de l'amélioration des pratiques enseignantes, l'évaluation des enseignements par les étudiants ne produit pas mécaniquement sur ces pratiques des effets bénéfiques et que l'apparition de ces derniers dépend des conditions de sa mise en œuvre. De nombreux travaux cherchent alors à identifier ces **conditions d'efficacité**. Elles peuvent être regroupées en cinq catégories.

Premièrement, le dispositif doit être conçu de telle sorte que les enseignants conservent la main mise sur le processus d'évaluation et ne sentent pas dépossédés de la gestion de leurs enseignements. Les enseignants doivent donc être invités à prendre part à l'élaboration collective du processus et ce, dès ses premières étapes : identification des fonctions de l'évaluation, discussion autour des critères, instauration des procédures et des modalités de circulation des informations, utilisations des résultats, types de retours et de suivis...

Deuxièmement, seule la fonction de contrôle requiert le recours systématique à un questionnaire standardisé et commun à tous les enseignants, quels que soient le cycle, l'année, la discipline et le programme. Si donc l'évaluation est authentiquement réalisée dans une perspective d'amélioration des pratiques, elle a tout intérêt à « coller » au plus près à ces pratiques et donc à s'adapter et à se contextualiser au maximum (e.g. possibilité d'ajouter ou de choisir les questions, présence de questions ouvertes...).

Troisièmement, de manière à atteindre des objectifs précis et différenciés, l'évaluation par les étudiants aura recours à une importante diversité de méthodes et de modalités, en veillant à chaque fois à leur adéquation par rapport à ces objectifs. Ainsi, un équilibre est à trouver entre des méthodologies quantitatives (e.g. questionnaires en ligne) et qualitatives (e.g. questions ouvertes, discussion entre l'enseignant et ses étudiants sur la base des résultats quantitatifs) ainsi qu'entre des procédures en cours de session (à visée formative) et en fin de session (à visée davantage sommative). Des outils très synthétiques peuvent être utilisés lorsqu'il s'agit d'un premier

repérage grossier d'éventuelles difficultés (e.g. la fonction de « détection de fumée institutionnelle » décrite en 2009 par Ricci), outils qui seront suivis par des analyses plus approfondies pour cerner précisément la nature et les causes des difficultés repérées.

Quatrièmement, de manière à favoriser leur implication, on veillera à informer clairement les étudiants des motifs, des procédures et surtout des effets de l'évaluation des enseignements, en mettant notamment en avant les améliorations des enseignements qui ont découlé des évaluations précédentes. Les étudiants seront impliqués dès l'amorce du processus et peuvent utilement être partie prenante des discussions portant sur les critères de qualité d'un enseignement (Gangloff-Ziegler & Weisser, 2011), dans une démarche d'éducation à la citoyenneté académique, telle qu'elle est prônée au chapitre 6.

Enfin et cinquièmement, bien qu'il s'agisse presque d'une lapalissade, un dispositif d'évaluation des enseignements ne produit des effets qu'à la condition qu'un accompagnement suive le recueil et la synthèse des données. Cet accompagnement va de l'aide à l'analyse des résultats, pour laquelle l'enseignant doit pouvoir bénéficier d'un appui professionnel *ad hoc*, à l'élaboration et à la mise en œuvre de pistes de solution face aux difficultés repérées, en passant par l'organisation des retours d'informations, tant à destination des étudiants que des équipes pédagogiques.

3.3.2 L'évaluation des programmes et des formations

Même si la distinction est finalement arbitraire et que les résultats de la première peuvent alimenter la seconde, l'évaluation des enseignements se distingue de l'évaluation de programme par le degré de granularité considéré. Traditionnellement, l'évaluation des enseignements par les étudiants porte sur des enseignements envisagés isolément, assurés par des personnes physiques, ce qui explique d'ailleurs que l'on peut glisser rapidement de l'évaluation de l'enseignement à celle de l'enseignant. Dans l'évaluation de programme au contraire (telle qu'elle est définie au chapitre 8 par Marthe Hurteau), c'est le programme dans son ensemble, en tant qu'assemblage d'activités de formation, qui fait l'objet de l'évaluation. L'unité d'analyse est bien l'ensemble de la formation dont les étudiants bénéficient et l'on cherche, par exemple, à en comprendre le fonctionnement (e.g. l'articulation et la cohérence en fonction des besoins ou des visées générales) et les effets (e.g. les compétences développées).

Malgré cette différence, les questions explorées par les études actuelles en matière d'évaluation de programme peuvent, comme celles qui ont trait à l'évaluation des enseignements, être regroupées en deux tendances principales : celles qui portent sur la validité et l'adéquation des méthodes, critères et procédures et celles qui étudient les conditions

favorables à la fois à la mise en place de telles évaluations et au fait qu'elles produisent les effets escomptés.

Le premier groupe de recherches aborde les questions de **validité** et d'**adéquation des méthodes, procédures et critères** de l'évaluation de programme en se demandant si les informations recueillies ainsi que les méthodes utilisées pour les traiter sont de nature à atteindre les objectifs visés, à savoir produire un jugement sur le programme et contribuer à en améliorer la qualité. Par exemple, on peut légitimement s'interroger sur la validité de ce que l'on appelle le « pilotage par les sorties ». Quelle est, en particulier, la validité d'un jugement porté sur un programme sur la base de mesures liées aux seules « sorties » du processus (e.g. le degré d'insertion professionnelle des diplômés), quand on sait que l'effet propre des établissements (à caractéristiques individuelles des diplômés et à conjoncture du marché du travail constantes) sur ce facteur est faible, voire non-significatif (Giret & Goudard, 2011) ?

Le même doute plane sur un critère pourtant fréquemment appelé à la rescousse, à savoir la satisfaction des étudiants. Ainsi, une étude montre que la corrélation entre la satisfaction rapportée par les étudiants par rapport à un programme et leur auto-évaluation des gains de compétences est assez faible (Schnoz-Schmied, Maupper & Balzer, 2011). Ce résultat indique que le développement de méthodes d'évaluation centrées sur les processus et les effets du programme notamment en termes d'acquis, au-delà de la seule satisfaction, est essentiel pour assurer la validité de l'évaluation de programme. C'est ainsi que sont actuellement investigués, en tant qu'indicateurs de qualité d'un programme, les intentions de transfert des étudiants, les transferts rapportés par ces derniers en situation professionnelle (ce qui passe par un suivi longitudinal comme le montre Trudel, 2011), les perceptions qu'ont les étudiants à l'issue d'un programme à propos de leurs gains d'employabilité ou de leur état de préparation à exercer une profession donnée (Limbach-Reich *et al.*, 2011).

Par ailleurs, des études montrent que la validité de l'évaluation de programme n'est assurée que par le recours à une importante diversité de méthodes, adaptées chacune aux visées de l'évaluation et aux critères retenus. Un bel exemple de cette diversité de méthodologies est fourni par Laurence Durat au chapitre 7. Elle montre bien que la poursuite d'objectifs d'évaluation diversifiés requiert la mise en place de méthodes multiples (allant de l'analyse des effets de la formation sur la carrière, la mobilité et le développement du réseau des formés jusqu'à l'identification des compétences perçues comme acquises et les évolutions identitaires) et complémentaires, tant quantitatives que qualitatives (e.g. entretien collectif de confrontation, dispositif de recueil de traces de l'activité en situation, méthode des incidents critiques).

Enfin, de manière à ne pas multiplier à l'excès les processus d'évaluation, des études s'interrogent sur la possibilité d'articuler l'évaluation de programme à d'autres types ou sources d'évaluation. C'est ainsi que Smidts (2011) a développé une démarche originale d'agrégation des résultats obtenus dans le cadre d'évaluations standardisées des enseignements par les étudiants à des fins d'analyse de la qualité d'un programme et de ses dynamiques internes.

Un second ensemble d'études s'intéressent, quant à elles, aux **conditions de mise en place** et aux **effets** de l'évaluation de programme. On y souligne notamment l'importance du suivi et du retour aux acteurs des résultats de l'évaluation, de la rigueur méthodologique de la démarche et, de manière plus générale, du fait que cette dernière aboutisse à un jugement perçu comme « crédible », c'est-à-dire satisfaisant en termes d'assurance quant à la solidité des procédures qui ont conduit à sa production et acceptable aux yeux des « détenteurs d'enjeux ». Le chapitre 8 est entièrement consacré à cet aspect délicat mais crucial de toute évaluation de programme : les bénéficiaires de l'évaluation, quels qu'ils soient, ne s'approprient ses résultats et ne les utilisent à des fins de régulation qu'à la condition d'être persuadés que le jugement produit rend effectivement compte, de manière juste et valide, de la réalité complexe du programme. Dans le cas contraire, ils ne se sentiront guère concernés par les aménagements que l'évaluation pourrait suggérer de mettre en œuvre sur la base des résultats obtenus.

Or, rendre compte de manière juste et crédible du fonctionnement et des effets d'un programme ou d'une formation n'est guère aisé. Par exemple, si l'on s'intéresse à la transférabilité, en situations professionnelles, des compétences acquises et que l'on cherche à savoir si le programme prépare effectivement à la gestion de ces situations, de redoutables questions méthodologiques ne manquent pas d'apparaître, comme celles évoquées par Laurence Durat au chapitre 7. Elle s'interroge d'abord sur l'écart, en partie irréductible, entre des programmes habituellement centrés sur l'acquisition de compétences individuelles et des exigences professionnelles supposant la mise en œuvre de compétences collectives, au sein de situations mouvantes et contingentes auxquelles la formation prépare finalement peu. Elle relève ensuite que les effets observés d'un programme peuvent ne pas être principalement ceux qui en étaient attendus, ce qui plaide pour une méthodologie d'évaluation très ouverte, permettant d'enregistrer des impacts imprévus, voire non-souhaités...

S'agissant des effets réels, des voix critiques s'élèvent de plus en plus fréquemment pour dénoncer le rapport « coût – efficacité » parfois peu convaincant de ces démarches très lourdes d'évaluation de programme. Ainsi, une étude extensive devrait être menée sur l'avenir que les responsables de programme réservent aux recommandations sur lesquelles ces

évaluations débouchent : on serait alors peut-être surpris de constater que la plupart des recommandations finissent par encombrer les étagères du bureau du responsable de programme sans qu'elles aient eu un quelconque impact sur ce dernier...

3.4 L'évaluation de la recherche et des chercheurs

S'il est une composante du métier d'enseignant-chercheur qui fait particulièrement débat de nos jours, c'est bien l'évaluation de la recherche et encore davantage celle des chercheurs. On le comprend aisément dans un monde universitaire où la qualité de chercheur constitue la valeur cardinale de réputation. L'ensemble des évolutions décrites au début de ce chapitre sont à l'œuvre dans ce domaine, de manière particulièrement vive. Ainsi, l'alourdissement considérable des tâches d'évaluation de la recherche, qui incombent aux enseignants-chercheurs dans un système fondé sur le *peer review*, est régulièrement dénoncé : évaluation des thèses de doctorat ; sélection des nouveaux enseignants et analyse des demandes de promotion ; évaluation de demandes de crédits de recherche, d'articles et d'ouvrages scientifiques ; compte rendu de publications scientifiques ; évaluation de candidats à des bourses ou postes de recherche ; évaluation de la recherche d'un laboratoire ou d'un établissement... Désirant par-dessus tout conserver la main mise sur les processus de financement et donc de reconnaissance de la recherche, les enseignants-chercheurs ont tendance à accepter ces nombreuses sollicitations d'évaluation, d'autant que leur multiplication constitue un signe de leur propre réputation de chercheur... Mais le maintien de ce pouvoir a un prix (Langfeldt & Kuvik, 2011) : 80 % des enseignants-chercheurs seraient impliqués dans de telles procédures qui leur demandent de plus en plus de temps (estimé à 20 jours par an en moyenne), notamment suite à l'externalisation du financement de la recherche et à la mise en compétition des équipes pour l'obtention de ce financement. De plus, des tensions naissent de cette surcharge des tâches d'évaluation (Langfeldt & Kuvik, 2011) : tension entre le fait de dégager du temps pour sa propre recherche et le temps nécessaire pour évaluer celle des autres, conflits d'intérêt dans un monde de la recherche largement mondialisé, tension entre neutralité et souci de promouvoir ses propres paradigmes, tension entre les jugements qualitatifs sur le contenu même de la recherche et l'usage d'indicateurs quantitatifs, tension entre autonomie et reddition de comptes...

Les travaux sur les processus d'évaluation de la recherche et des chercheurs sont donc nombreux et le colloque de l'ADMEE a montré qu'ils peuvent être regroupés en deux principales tendances : d'une part, l'étude de la validité des procédures et des critères des évaluations actuelles (sont-elles réellement à même de mesurer la qualité des recherches produites ?) et, d'autre part, l'étude des effets, notamment sur les pratiques elles-mêmes

de recherche, de la manière dont cette dernière est appréciée et mesurée (dans quels sens l'évaluation, en tant que regard interprétatif porté sur un objet, contribue-t-elle à transformer cet objet, en l'occurrence les pratiques elles-mêmes de recherche ?).

3.4.1 La validité des procédures et des critères

Si les procédures classiques d'évaluation de la recherche par le système de *peer review* ne sont pas exemptes de défauts (lourdeur ; lenteur ; risque de copinage, d'endogamie scientifique et de reproduction des hiérarchies existantes...), les modalités davantage quantitatives et formalisées qui se sont développées ces dernières décennies et qui ont recours aux indicateurs bibliométriques (e.g. l'indice *h*) posent de redoutables problèmes de validité. De nombreuses études se sont attelées à identifier les problèmes essentiellement de validité et donc de pertinence de ces nouvelles façons d'évaluer la recherche. À tel point que l'on peut raisonnablement s'interroger sur le bien-fondé de la suprématie actuelle des procédures standardisées basées sur des quantifications externes au détriment des jugements produits par des pairs à propos de l'objet même des recherches. Aux chapitres 9 et 10, Jean-Marie De Ketele et Rémi Goasdoué montrent d'ailleurs que les critères de qualité d'une recherche sont variés et multidimensionnels et sont donc très peu susceptibles d'être ramassés, de manière valide, dans des nombres et *a fortiori* dans un seul...

Sans revenir sur les nombreuses anomalies qui ont été utilement mises en évidence en matière d'évaluation quantitative et automatisée de la recherche, citons quelques études critiques qui dénoncent le manque de validité de ces procédures actuellement très en vogue. Rappelons d'abord qu'elles se fondent de plus en plus souvent sur des indicateurs bibliométriques censés rendre compte de la qualité de l'activité des chercheurs via leurs publications. Ces indicateurs ont recours à de multiples comptages en cascade : pour apprécier la « productivité » des chercheurs, il faut, par exemple, mesurer la quantité et la qualité de leurs publications, ce qui suppose de distinguer la qualité des revues dans lesquelles ils publient et donc leur facteur d'impact. Ces différents comptages donnent l'illusion de l'objectivité, alors qu'ils reposent sur de nombreux choix techniques et méthodologiques, qui ne sont bien sûr jamais neutres en termes de valeurs et qui parfois mettent à mal la validité du jugement final. À titre d'exemple, Méla (2008) a montré que le facteur d'impact de journaux de mathématiques était calculé à partir d'une base de données dans laquelle n'était répertoriée que la moitié des journaux référencés dans les deux principaux *reviewing journals* de la discipline et que les citations prises en compte ne concernaient que les deux dernières années alors que, dans la base de donnée principale de la discipline, 90 % des articles sont cités en dehors de cette période de deux ans.

Gingras (2008) a par ailleurs montré que les indicateurs bibliométriques sont loin de posséder deux des propriétés essentielles dont devrait se prévaloir, surtout s'il prétend participer à une évaluation individuelle des chercheurs, tout « bon » indicateur : l'adéquation à l'objet (e.g. le seul indice des citations peut-il prétendre fournir une image correcte de la qualité de la recherche ?) et l'homogénéité de la mesure (e.g. le nombre d'articles publiés combiné à un indice de réputation ne constitue-t-il pas un indicateur très hétérogène sans signification certaine ?).

À côté des problèmes techniques, statistiques et culturels[110] qu'ils soulèvent, on peut plus fondamentalement s'interroger sur le sens et l'adéquation de ces mesures et indicateurs. Peut-on *mesurer* la production scientifique et la qualité d'une recherche ? De manière directe, l'opération semble délicate en raison de l'absence d'une grandeur unique de référence et de la multidimensionnalité de l'objet à mesurer (Blay, 2009). On se rabat alors sur des indicateurs indirects : le nombre de publications, de citations... Mais sont-ils eux-mêmes valides, c'est-à-dire permettent-il de dire quelque-chose d'approximativement vrai sur l'objet à évaluer, en l'occurrence la qualité de la recherche ? Par exemple, que signifie *l'impact* d'un chercheur et quelle fonction remplit une *citation ?* Un taux de citations élevé signifie bien sûr que des chercheurs se sont appuyés d'une certaine manière sur l'auteur cité, ce qui pourrait en effet laisser sous-entendre que sa recherche a été d'une certaine qualité. Mais s'agit-il de citation de réelle dette intellectuelle ou de citation remplissant en réalité une autre fonction, rhétorique (e.g. choisie en fonction du prestige de l'auteur ou de sa proximité) ou d'allégeance stratégique ? On en oublierait que la meilleure manière de se faire un avis sur la qualité d'une recherche est d'en lire de manière critique et exhaustive le compte rendu...

3.4.2 Les effets de la mesure sur l'objet évalué

Grand classique de la docimologie, un déficit de validité entraîne bien souvent un problème de fidélité : quand on ne sait pas trop ce que l'on évalue, on risque d'aboutir à des mesures peu stabilisées, que ce soit dans le temps ou selon les outils utilisés. Ainsi, Milard (2010) a montré que les indices de citations ne mettaient pas en évidence les mêmes auteurs selon l'outil utilisé (l'un était plutôt centré sur la littérature académique et l'autre sur la présence des auteurs sur Internet). Comment accorder du crédit à des indicateurs dont les mesures aboutissent à des résultats très différents selon l'outil envisagé ?

110. Dans certaines disciplines en tout cas, l'hégémonie anglo-américaine assurée par le fonctionement même des méthodes de bibliométrie aboutit aussi à un problème de validité : ce que l'on mesure n'est plus la qualité de la recherche produite mais sa proximité avec les cadres de référence et les paradigmes de la recherche anglo-saxonne (Beauvois, 2005).

Les effets les plus délétères ont sans doute trait aux modifications des habitudes de publication et des pratiques elles-mêmes de recherche qu'entraîne l'usage intensif des indicateurs bibliométrique pour évaluer la recherche. On s'aperçoit que la façon de mesurer la qualité de la recherche finit par affecter la manière dont les chercheurs se conduisent, et pas toujours au bénéfice d'une plus grande qualité de leurs travaux. Ainsi, se multiplient des pratiques d'écriture douteuses : signature systématique d'articles par l'entièreté du laboratoire, auto-citations intempestives, citations de complaisance de chercheurs encore en activité ou du moins en vie et qui peuvent rendre la pareille, émiettement du compte-rendu d'une recherche sur plusieurs papiers, expression d'opinions paradoxales dont le seul but est d'attirer la réplique (et donc d'être cité).

Plus fondamentalement, les modalités de mesure de la qualité de la recherche ont pour effet de réorienter les pratiques elles-mêmes de recherche en fonction de ce qui sera aisément publiable et de ce qui assurera un impact fort. Le plus grand risque est en définitive de nuire à la créativité, pourtant essentielle en matière de production de savoirs nouveaux, les équipes privilégiant des thématiques et des approches standardisées, reconnues et consensuelles. Tout ce passe comme si une étrange inversion s'était produite. Alors qu'auparavant la recherche et ses enjeux étaient les premiers éléments de réflexion du chercheur qui se demandait ensuite, une fois les résultats engrangés, comme les faire connaître via une publication, ce sont actuellement les contraintes qui pèsent sur les modes de publication les plus « porteurs » en termes de citation et de réputation qui déterminent en amont le contenu et les modalités de la recherche à mener.

À terme, il n'est pas déraisonnable de penser que ces nouvelles modalités de mesure de la recherche produiront des changements identitaires chez les chercheurs et des modifications organisationnelles, voire de valeurs au sein de leur communauté, les règles de fonctionnement se calquant de plus en plus sur l'outil de mesure, dans une logique du *game playing* où l'essentiel est d'obtenir les résultats les plus performants (Milard, 2010). Ainsi, un des risques majeurs à long terme réside dans une standardisation des recherches, en faveur de celles qui seront anticipativement jugées dignes d'être poursuivies, en fonction de leurs possibilités de publication, voire de leurs retombées commerciales dans la logique d'une économie du savoir.

On notera enfin que les effets économiques ne sont pas négligeables et auront à terme des conséquences sur les règles de fonctionnement du monde de la recherche. Ainsi, on assiste à la création d'un « marché » de l'évaluation de la recherche dans lequel agissent désormais de nombreuses entreprises privées (propriétaires de revue, agences de citations...). En matière de diffusion, les nouvelles procédures ont aussi des conséquences

sur les maisons d'édition et notamment sur leur standardisation consécutive à la disparition progressive des revues non cotées (Milard, 2010).

3.5 L'évaluation institutionnelle

On regroupera ici les évaluations qui portent sur des entités considérées comme des unités fonctionnelles (laboratoires, départements, facultés et établissements) et dont on cherche à mesurer la qualité et/ou la performance. Ces évaluations se réalisent principalement selon deux perspectives. Il s'agit d'abord, dans une logique de reddition de comptes, de s'interroger sur le bon usage que ces entités font du financement public. Mais cette préoccupation en arrive vite à une autre qui consiste à classer ces entités les unes par rapport aux autres, dans le but soit de les mettre en compétition face à un financement différencié (e.g. financer davantage les laboratoires « publiants ») soit d'informer les clientèles, comme on dit désormais, de leur supposée qualité relative (e.g. pour guider le choix de partenaires de recherche[111] ou éclairer les enseignants et les étudiants dans un contexte de mobilité internationale croissante).

Si la nécessité des formes d'évaluation évoquées jusqu'ici n'était que rarement remise en cause, il en va tout autrement de l'évaluation institutionnelle. Beaucoup dénoncent le développement exponentiel et par addition d'une jungle nationale et internationale de procédures et de modalités d'évaluation institutionnelle, consécutif à l'apparition de deux phénomènes : la création d'agences nationales de la qualité (notamment en Europe dans le cadre du processus de Bologne (e.g. l'AERES décrite et analysée au chapitre 12) et le développement de « palmarès internationaux » portant sur les établissements.

Pressentant qu'un jour ou l'autre un lien explicite sera réalisé entre ces évaluations institutionnelles et les décisions de financement, les acteurs de l'enseignement supérieur se montrent particulièrement attentifs à la validité et la solidité des procédures et des indicateurs auxquels elles ont recours. Et dès que l'on creuse un peu cette question, de sérieux doutes ne manquent pas de s'installer...

3.5.1 Les évaluations nationales de la « qualité »

S'agissant de l'évaluation institutionnelle de la qualité de la formation[112], la question la plus épineuse est de savoir comment mesurer

111. Il s'agissait d'ailleurs à l'origine du but premier du classement de Shanghai.
112. L'évaluation institutionnelle comprend des aspects très divers, dont certains ont trait à la gouvernance et à la gestion des personnels et des budgets, notamment dans les pays qui, comme la France, connaissent une autonomie croissante de leurs établissements d'enseignement supérieur. Ces aspects ne seront pas abordés dans ce chapitre conclusif, portant uniquement sur l'évaluation des missions de formation et de recherche.

la qualité en cette matière, en l'absence de possibilité d'apprécier directement les « produits » des formations, à savoir les acquis des étudiants. En effet, l'inexistence d'épreuves standardisées nationales ou internationales portant sur les acquis des étudiants contraint ce type d'évaluation à s'en remettre à l'analyse des processus uniquement : les flux d'étudiants et leurs taux de réussite, l'attractivité des filières, des diplômes et des établissements (notamment vis-à-vis des étudiants étrangers), les taux d'insertion des diplômés, le respect des normes de qualité de formation (e.g. charte de l'évaluation, référentiel de compétences, descriptif des cours). Mais tout cela ne garantit pas encore que la formation soit de qualité : un « bon » taux de réussite peut résulter d'un laxisme dans la certification, l'attractivité est davantage liée à la réputation, l'insertion professionnelle des étudiants dépend surtout de la spécialisation qu'ils ont suivie (et de son adéquation au marché du travail) et des disparités régionales de ce marché (Giret & Goudard, 2011). De ces limites, découle le sentiment actuel de la nécessité de développer de nouveaux indicateurs (e.g. le « bien-être » des formés et leur sentiment d'appartenance à l'établissement).

Une autre façon de contourner le problème de l'absence de mesures directes est de s'en remettre à l'auto-évaluation : sera considéré comme de qualité et performant l'établissement (ou la filière) qui, au terme d'une introspection outillée, estime qu'il arrive à ses fins. D'où le succès grandissant, dans les procédures d'évaluation, de la phase dite d'auto-évaluation, dont certains redoutent qu'elle soit devenue la « version sécularisée de la bonne vieille confession » ou « la généralisation de l'auto-critique maoïste » (Büttgen & Cassin, 2009).

On voit aussi comment cette forme d'évaluation témoigne de l'apparition d'un nouveau type de relation entre l'État et les établissements d'enseignement supérieur, comme le montrent David Carassus et Élodie Dupuy au chapitre 13. Dans la perspective du *new public management*, l'État ne souhaite plus attribuer indistinctement son financement sur la base d'indicateurs administratifs et neutres (e.g. le nombre d'étudiants). Il désire au contraire différencier le financement accordé aux filières et aux établissements sur la base de leurs « résultats » et de leurs « performance » qu'il s'agit alors de mesurer à partir d'indicateurs souvent indirects. L'État se transforme ainsi en État-évaluateur. Cette évolution est particulièrement perceptible dans la gestion d'une partie du « Grand Emprunt » français évoqué au chapitre 12, que ce soit à propos des IDEX (Initiatives d'Excellence) ou des IDEFI (Initiatives d'Excellence en Formations Innovantes). Ces financements impulsionnels n'ont pas été accordés *a priori* aux établissements, mais ils ont été octroyés sur la base de projets mis en compétition les uns avec les autres et donc évalués par une agence (l'ANR), faisant elle-même appel à un panel d'experts. Le processus est assez lourd, chronophage

à la fois pour les établissements qui déposent des projets et pour l'agence et ses évaluateurs, sans que l'on soit tout à fait sûr que l'efficacité des projets mis en place en définitive soit largement supérieure à celle d'un système dans lequel les établissements auraient reçu d'office un financement à consacrer aux innovations et auraient été invités à rendre des comptes de leur utilisation.

D'autres études développent des approches comparées des outils et des procédures des évaluations nationales, comme la recherche EVALUE présentée au chapitre 14. On s'intéresse également aux conditions d'efficacité de ces évaluations. Macarie-Floréa (2010) a ainsi analysé le fonctionnement du Conseil National d'Évaluation français en cherchant à savoir à quelles conditions les évaluations institutionnelles (et en particulier les phases d'auto-évaluation) avaient eu de réels effets auprès des établissements. L'auteur souligne que l'implication forte et volontariste des équipes de direction et la préexistence d'une culture de l'évaluation constituent deux facteurs d'efficacité majeurs.

3.5.2 Les « hit-parades » internationaux

S'agissant des classements internationaux des établissements, les critiques sont encore plus acerbes. Cette étrange frénésie des classements et des palmarès internationaux a envahi le monde de l'éducation et de la formation à partir des années 1990. Un syndrome « eurovision » a progressivement contaminé la réflexion sur les systèmes éducatifs, sans doute par la facilité et le pouvoir médiatique fort du principe même du hit-parade. Par exemple, le rang des pays au sein des classements PISA[113] semble être devenu, pour l'enseignement secondaire, l'unité d'une échelle de mesure de la qualité de leurs systèmes éducatifs, comme le degré Celsius l'est à la température : on dira laconiquement que l'enseignement de tel pays n'est pas « efficace » car il se classe à la 20e position (sur 30) du dernier palmarès PISA... Malgré les (ou peut-être aussi à cause des...) trop nombreux raccourcis qu'implique à l'évidence un pareil jugement (Romainville, 2002), ces classements ont connu un succès grandissant sans doute parce qu'ils constituent à la fois un indice et un vecteur de la mondialisation (Lussault, 2010) ainsi qu'un outil de positionnement dans la grande « bataille » de la compétition internationale pour l'excellence que cette mondialisation a engendrée (Hazelkorn, 2011).

Le premier coup de semonce est venu de la publication, en 2003, du classement de l'université Jiao Tong de Shanghai (*Academic Ranking*

113. Géré par l'OCDE, le programme PISA (Programme International pour le Suivi des Acquis des élèves) est un ensemble d'études visant à mesurer les « performances » des systèmes éducatifs des pays membres et non membres, sur la base de comparaisons entre les acquis des élèves d'un âge donné.

of World Universities) qui range les universités[114] de la planète (en général, par domaines et par disciplines scientifiques) selon essentiellement des critères de réputation en recherche. Par exemple, 30 % de la note de l'établissement dépend du nombre de prix Nobel ou de médailles Fields obtenus par ses professeurs ou ses diplômés. D'autres classements ont suivi, produits par des organismes privés ou publics. Un certain nombre d'entre eux ont été élaborés ou suscités par des organismes, des pays ou des établissements qui estimaient être mal classés injustement par Shanghai... Et c'est ainsi qu'un des premiers reproches que l'on peut adresser à ces classements réside dans le manque de stabilité de leurs résultats. S'ils prétendent mesurer la qualité intrinsèque d'un établissement, on ne voit pas pourquoi l'ordre serait bouleversé suivant l'instrument de mesure. Or, dans le haut du classement du très britannique *Times Higher Education*, on observe une large surreprésentation des universités anglaises et des universités issues de l'ex-Empire. Le classement de Leyden reprend plusieurs universités hollandaises dans son « top 15 », celui de l'École des Mines classe assez avantageusement quelques Grandes Écoles françaises...

De nombreuses études ont été consacrées à **la validité et la fiabilité de ces classements**, avec des résultats souvent très critiques (Bellon, 2007 ; Gingras, 2008 ; Hazelkorn, 2011 ; Matzkin, 2009 ; Salmi & Saroyan, 2007) :

- grande difficulté de réplication en l'absence d'explicitation des procédures et d'accès public aux bases de données constituées (surtout le classement de Shanghai) ;
- trop grande sensibilité de l'instrument à certaines données isolées et aux changements de méthodologie (e.g. l'obtention d'un Nobel peut faire gagner 100 rangs) ;
- hétérogénéité des indicateurs qui ne devrait pas autoriser la constitution, via une pondération arbitraire, d'une note globale et plus fondamentalement inadéquation d'un score et d'une méthode uniques en regard de la grande diversité des missions poursuivies par tout établissement d'enseignement supérieur ;
- accent exagérément placé sur le classement : des différences de places peuvent ne pas être significatives ou ne sont que marginalement significatives alors qu'elles vont engendrer des écarts importants de réputation et ultérieurement de ressources ;
- absence de prise en compte du type et de la taille des établissements ;
- recours trop massif à l'avis subjectif de pairs (e.g. sur la réputation en enseignement) pour lequel on peut redouter l'existence d'un

114. En réalité, un tout petit pourcentage d'entre elles seulement : 500 environ sur plus de 15 000 existantes...

effet de halo : on aura en effet tendance, sur la base d'une appréciation très positive d'une dimension réputée d'un établissement, à rendre plus positive l'évaluation d'autres dimensions, même sans les connaître ;
- surclassement des établissements anglophones ou enseignant en anglais, des établissements répondant au modèle américain des universités de recherche et au modèle des établissements privés et sélectifs (tant en ce qui concerne les étudiants que les enseignants) capables de lever des fonds privés de recherche via notamment des fondations ;
- quasi absence d'influence de la production scientifique en arts, lettres et sciences humaines en raison de leur faible impact bibliométrique ;
- évaluation totalement décontextualisée : appréciation de l'ensemble des établissements selon un petit nombre de critères identiques quel que soit leur profil de missions notamment[115] et absence de prise en compte des cadres et modes de financement nationaux et du contexte dans lequel s'insèrent les établissements. Ainsi, alors que les États-Unis arrivent largement en tête de la plupart des palmarès des établissements par pays, ils dégringolent de plus de dix places si les classements sont réalisés en tenant compte du nombre d'habitants ou du produit intérieur brut (Hazelkorn, 2011).

D'autres études ont été consacrées aux **effets délétères du prurit des classements internationaux**. Évaluer revient à apprécier la valeur d'une chose en fonction de critères. L'évaluation ne se réduit cependant pas à la production d'un jugement de valeur sur cette chose. Elle induit aussi, en amont, une réorientation des pratiques qui vont être davantage centrées sur ce qui aura été considéré comme de plus grande valeur. Et c'est particulièrement vrai pour les palmarès : classer revient toujours à inviter les derniers à produire des efforts pour gagner quelques places au prochain classement et donc à se rapprocher des modes de fonctionnement des « mieux classés ». Le jeu est par ailleurs sans fin, puisque c'est dans la nature même du classement qu'il y ait des derniers et des premiers...

L'on voit bien que les effets en retour des classements internationaux auprès des dirigeants (ministres, administration, présidents d'établissements) risquent d'être bien réels puisque ceux-ci auront tendance à axer leur politique sur « ce qui compte » dans ces classements. On peut d'ailleurs se demander si la politique très en vogue de regroupements frénétiques d'établissements n'a pas constitué un des premiers dommages collatéraux des classements internationaux, tout comme la volonté qu'expriment de

115. Ce qui revient, comme le note avec humour Hazelkorn (2011), à ranger les établissements selon leur écart par rapport aux caractéristiques d'Harvard...

plus en plus de pays de privilégier quelques établissements « d'excellence », susceptibles d'apparaître dans le sommet des palmarès.

L'**effet d'image** est loin d'être négligeable : malgré leurs lacunes méthodologiques évoquées ci-dessus, les palmarès ramassent dans une note, une position ou une place dans le « top x » l'ensemble des activités d'un établissement. Bien qu'elle soit largement réductrice et qu'elle devrait être remise en contexte et interprétée avec prudence, cette position fait ensuite l'objet d'une médiatisation outrancière. S'il est sans doute peu contestable que les premiers classés soient très performants, il est beaucoup plus discutable de prétendre que les établissements « mal classés » et surtout les « non-classés » (la grande majorité des universités, par exemple, dans le classement de Shanghai) présentent une performance médiocre (Lussault, 2010). C'est pourtant de cette manière qu'ils seront désormais perçus par le grand public.

Plus fondamentalement, on peut craindre une **uniformisation des systèmes universitaires et de leur mode d'organisation** sur le modèle des établissements « bien classés ». Les responsables universitaires pourraient avoir tendance à désormais organiser leurs établissements selon les règles qui ont valu aux « mieux classés » leur position enviable. À l'évidence et compte tenu de la diversité des fonctions et des contextes des établissements, il s'agirait d'un appauvrissement considérable de l'apport de l'enseignement supérieur à la société (Lussault, 2010).

Les critères retenus et notamment la prédominance des critères bibliométriques aboutissent également à la **dévalorisation de certaines disciplines** qui interviennent peu dans les classements (lettres et sciences humaines) à cause du type de revues prises en compte (e.g. les incontournables *Nature* et *Science*, miroirs des sciences dites « exactes » et « à laboratoires »).

Plus grave encore, les palmarès étant organisés à l'aulne de critères se prêtant aisément à la quantification automatique, on pourrait assister à une **dévalorisation, voire à l'abandon de missions peu visibles et peu mesurables**. Ce serait notamment le cas, comme si elles ne souffraient déjà pas d'un manque d'estime suffisant, des missions enseignement. La qualité de l'enseignement est en effet peu prise en compte par les classements et lorsqu'elle l'est, c'est uniquement sur la base de critères réputationnels (e.g. enquête de réputation sur l'enseignement pour le classement du *Times* comptant pour 15 %) ou de critères dont on peut douter du lien avec la qualité de l'enseignement (e.g. le revenu des enseignants pour 2,25 % dans le même classement). Mais cette dévalorisation risque de toucher un ensemble plus large de missions de l'enseignement supérieur qui ne se prêtent pas à une comptabilisation aisée et automatique, même si elles peuvent se révéler cruciales pour la société : missions de service

(en particulier régional ou local), participation à la création et à l'innovation sociales et culturelles, diffusion de l'innovation au-delà des revues scientifiques et notamment dans le tissu économique...

4. IMPLICATIONS POUR LES PRATIQUES : DES ÉQUILIBRES À RÉINVENTER

Au final, l'image générale des relations qui unissent actuellement l'enseignement supérieur à l'évaluation est plutôt celle d'une série de malentendus, de discordes et de tensions, qui portent assez rarement sur le bien-fondé même de l'évaluation, mais qui a davantage trait aux modalités selon lesquelles elle a tendance à être organisée ces dernières années. Si l'on souhaite réconcilier les acteurs de l'enseignement supérieur avec une évaluation trop souvent perçue aujourd'hui comme frénétique et intrusive, un certain nombre d'équilibres sont à réinventer au sein même des pratiques évaluatives. C'est par un inventaire de ces arbitrages aujourd'hui nécessaires à une réhabilitation de l'évaluation dans l'enseignement supérieur que la présente conclusion se terminera.

4.1 Trop d'évaluation tue l'évaluation

L'évaluation finissant toujours par s'imposer ne fût-ce que sauvagement, la question n'est pas de savoir s'il convient ou non d'évaluer, mais plutôt de choisir entre deux scénarios : soit laisser se développer une évaluation inévitable, sauvage et hors de contrôle, soit organiser une évaluation instituée selon des règles et des procédures explicites. Dans tous les domaines, l'évaluation instituée a l'avantage d'être plus transparente et plus démocratique. Elle permet de réduire l'opacité et ainsi de combattre l'arbitraire, le copinage et l'injustice. Mais elle est également très lourde à mettre en œuvre et consomme beaucoup de temps et d'énergie à la fois auprès des personnes qui la mettent en œuvre mais aussi auprès de ceux qui en font l'objet. Il suffit pour s'en convaincre de penser aux procédures parfois très invasives d'évaluation des enseignements par les étudiants qui se sont développées ces dernières années et dont les enseignants et les étudiants dénoncent la lourdeur.

En matière d'évaluation aussi, l'enfer est pavé de bonnes intentions et le mieux est l'ennemi du bien : face aux dangers et dérives de l'évaluation spontanée ou implicite, on en fait vite un peu trop en faveur de son institutionnalisation. Compte tenu de la complexité des objets à évaluer, le processus s'engage alors dans une direction inflationniste à la fois néfaste et illusoire. Néfaste, car les acteurs impliqués dans une forme instituée d'évaluation se montrent rapidement épuisés par son technicisme et son formalisme, ce qui peut les conduire à rejeter un processus qui se serait trop emballé à leurs yeux. Illusoire, car le mirage scientiste et positiviste qui peut

toujours présider à pareille entreprise fait croire que l'évaluation a été à ce point rendue transparente et systématique qu'elle en serait devenue objective et incontestable, ce qui n'est jamais exact.

Un premier équilibre consiste donc à trouver une juste mesure entre *trop peu* et *trop* d'évaluation instituée. Il convient certes de baliser fortement le processus d'évaluation en déterminant des procédures claires et transparentes, mais celles-ci doivent également veiller à rester aussi souples et légères que possible. À titre d'exemple, cet équilibre est actuellement souhaitable dans le domaine de l'évaluation des enseignements par les étudiants. Certes l'instauration de modalités explicites dans ce domaine permet d'éviter les jugements de valeur intempestifs et les règlements de compte internes réalisés sur la base d'ouï-dire ou de bruits de couloir. Cependant, il est aussi contre-productif de se lancer dans des procédures trop complexes (questionnaires trop lourds), longues (appliqués à l'ensemble des enseignements) et répétitives (chaque année), qui épuiseront rapidement les acteurs au point qu'elles seront finalement abandonnées après quelques années.

4.2 La mesure (et encore moins le chiffre...) ne fait pas l'évaluation

À plusieurs reprises, a été dénoncée dans cet ouvrage la « religion du chiffre » qui s'est installée dans l'enseignement supérieur en matière d'évaluation. On n'hésite plus à ramener la qualité d'un établissement à son rang dans un classement international, celle d'une revue à une lettre (A, B, C). Malgré le caractère évidemment abusif et pour le moins réducteur de ces raccourcis, certains d'entre eux finissent par avoir des effets importants, comme par exemple la distinction, encore plus sommaire puisque binaire, entre les chercheurs « publiants » et « non-publiants ».

Une des critiques les plus transversales aux chapitres précédents a trait au réductionnisme auquel l'évaluation dans l'enseignement supérieur semble avoir cédé. Pressé par la nécessité de standardiser l'évaluation, on en est malheureusement arrivé à ramener des réalités humaines et sociales, par essence complexes, à quelques chiffres ou lettres censés les mesurer. Or la mesure n'est pas l'évaluation : si l'on se réfère à la définition proposée ci-dessus, la mesure ne constitue qu'une phase du processus d'évaluation, importante sans doute en matière de recueil systématique d'informations, mais qu'il faut encore interpréter et analyser pour élaborer un jugement.

On pourrait s'interroger longuement sur cette étrange fascination actuelle autour du chiffre et autour des procédures automatisées, lisses et d'apparence objective qui le produisent. Antoine Prost avance une hypothèse intéressante en conclusion de l'ouvrage dirigé par Emin et Villeneuve (2009). Pour lui, ce repli vers le chiffre serait le signe d'un changement important du rapport à l'autorité. Considérant désormais que l'imposition

hiérarchique de jugement de valeur est socialement inacceptable, on cherche à élaborer d'autres modalités d'imposition moins suspectes d'autoritarisme et apparemment plus neutres parce que censées résulter de processus objectifs. Par là même, la religion du chiffre ne serait rien d'autre qu'un refuge face à la peur de fonder clairement son jugement sur des valeurs dont on n'oserait plus se réclamer. L'expert externe s'est substitué au supérieur hiérarchique. Or, nous y reviendrons au point suivant, toute évaluation est fondamentalement une opération qui consiste à déterminer la valeur d'une chose et il est donc impossible de se réfugier dans l'externalité impartiale de la seule mesure.

Une autre explication plus prosaïque est que la mesure constitue en réalité la phase la plus aisée et la plus tranquille de l'évaluation : beaucoup de choses se laissent compter et une sorte de paresse intellectuelle fait que l'on se concentre alors sur le détail des mécanismes de comptage davantage que sur la question centrale, qui est de savoir si ce que l'on compte est bien en rapport avec ce que l'on souhaite évaluer. Comme l'écrivaient avec humour Casey et ses collègues (1997, p. 474), « il n'est généralement pas trop difficile de mesurer quelque chose dans le domaine de l'éducation, la difficulté est plutôt de savoir ce que l'on mesure exactement »... Et le problème n'est pas neuf : dès 1921, un célèbre psychologue mettait en garde ses collègues quant à la démangeaison irrationnelle de mesure qui menaçait alors la psychologie « scientifique » naissante (cité par Blay, 2009, p. 20) :

> N'ayant pas le moyen de mesurer ce que vous désirez, la démangeaison de mesurer peut, par exemple, simplement vous conduire à mesurer quelque chose d'autre – et peut-être en oubliant la différence – ou en laissant de côté certaines choses parce qu'elles ne sont pas mesurables.

Ici aussi, les pratiques évaluatives devraient sans doute tendre vers la recherche d'un meilleur équilibre entre, d'une part, la nécessité de recourir à des procédures standardisées de recueil d'informations (incluant des opérations de mesure adéquates) assurant à la production du jugement évaluatif sa solidité et, d'autre part, le respect de la complexité de la chose évaluée, ce qui suppose de conserver une place importante au débat, à l'argumentation et à l'appréciation qualitative.

Ainsi, l'évaluation de la recherche devait sans doute en revenir davantage à l'avis argumenté de pairs sur le contenu même des recherches produites, tout comme l'évaluation des acquis des étudiants, surtout dans une approche par compétences, devrait sans doute abandonner l'illusoire échelle de notes sur 20 points pour se tourner vers des diagnostics plus fins portant sur les composantes de la compétence acquises ou non par l'étudiant.

Par ailleurs, l'évaluation n'a pas à recourir, systématiquement et dans tous les cas de figure, au chiffre sous peine d'une dangereuse dérive qui consiste à renoncer finalement à évaluer ce qui semble délicat à mesurer, selon la célèbre maxime qu'Albert Einstein avait affichée dans son bureau de Princeton : « Ce qui compte ne peut pas toujours être compté, et ce qui peut être compté ne compte pas forcément ».

4.3 Évaluation sans valeur n'est que ruine de l'âme

Dès lors qu'évaluer revient toujours à apprécier la valeur d'une chose, il n'y a pas d'évaluation sans échelle de valeurs. On évalue toujours en regard de ce que l'on considère, explicitement ou implicitement, comme souhaitable. Et l'une des qualités de l'évaluation est d'ailleurs qu'elle se déroule en regard de valeurs explicitement énoncées au préalable, ce qui n'est malheureusement pas toujours le cas. Par exemple, évaluer la qualité d'un établissement par sa position au sein d'un classement international n'est pas neutre : l'évaluation se déploie alors implicitement dans une approche « client », selon une logique qui consiste à évaluer la qualité des services rendus à ces clients. C'est une option. Une autre serait, dans une approche « bien public », d'évaluer des paramètres différents comme les conditions d'accès du plus grand nombre aux services rendus, ce qui peut conduire – à l'évidence – à deux appréciations discordantes (Lussault, 2010).

L'évaluation est d'autant moins neutre en termes idéologiques qu'elle ne se réduit pas à apprécier *a postiori* la valeur d'un objet : elle oriente aussi, en amont, les pratiques de production de cet objet. Lorsqu'une procédure d'évaluation adopte une liste de critères communs et prétendument universels, comme le font les classements internationaux des établissements, elle induit par là des conduites communes et une standardisation des pratiques (Bellon, 2007) : les établissements sont amenés à réorienter leurs modes de fonctionnement selon des normes qui leur assureront une chance de « monter » dans ces classements. À l'opposé, si l'on estime, selon un autre modèle, qu'évaluer des établissements revient à apprécier l'atteinte de leurs objectifs, l'évaluation sera axée sur une meilleure connaissance des objectifs que chaque établissement s'est fixés et sur l'examen de la cohérence entre ces objectifs, les moyens mis en œuvre et les réalisations : tout le contraire des classements internationaux...

Dans le même sens, le chapitre 13 montre bien que l'évaluation de la performance globale des établissements conduit à porter sur ces derniers un nouveau regard et à les considérer comme des unités fonctionnelles, autonomes et susceptibles d'avoir une politique institutionnelle forte. Il en va de même pour l'évaluation des acquis des étudiants : ces derniers se conforment, dans leurs manières d'apprendre, à la façon dont

l'évaluation sera organisée et aux aspects sur lesquels elle portera (cf. chapitre 1). L'évaluation modifie donc l'objet auquel elle s'applique et finit par rendre réels et effectifs les critères auxquels elle recourt et les valeurs sur lesquelles elle se fonde : « Quand les hommes définissent des situations comme réelles, elles sont réelles dans leurs conséquences » (cité par Salmi & Saroyan, 2007, p. 34).

Zarka (2009) émet d'ailleurs une hypothèse intéressante selon laquelle le succès du recours massif aux procédures standardisées et d'apparence neutre s'explique par le fait que ces procédures masquent en réalité les conflits et les débats de valeur inhérents à toute évaluation. Elles nous offriraient le confort illusoire de l'objectivité externe, comme si celle-ci pouvait être fondée sur des valeurs indiscutables, voire se passer même d'un débat sur les valeurs.

Il faut donc réaffirmer qu'aucune mesure et qu'aucun indicateur ne sont neutres et qu'ils n'ont pas la capacité, à eux seuls, d'entraîner « objectivement » un jugement et encore moins une décision. Par exemple, une procédure, qui elle-même repose déjà sur des choix de valeurs tels que le type de productions et de revues à prendre en compte, peut aboutir à isoler les « publiants » des « non-publiants ». Mais cette procédure ne dit encore rien de la décision à prendre ensuite : s'agit-il d'attribuer aux « publiants » davantage de crédits de recherche parce qu'ils les mériteraient ou, au contraire, dans une sorte de politique de discrimination positive de la publication, s'agit-il d'accorder plus de moyens aux « non-publiants » et notamment aux jeunes de manière à les inciter et à les aider à diffuser les résultats de leurs recherches ?

Il convient donc de replacer les valeurs au centre des pratiques d'évaluation. L'équilibre à trouver consiste ici à se focaliser autant sur la fiabilité des procédures et des modalités que sur l'explicitation et la discussion collectives des valeurs sur lesquelles elles se fondent. Par exemple, les critères affleurant dans les questionnaires d'évaluation de l'enseignement par les étudiants devraient faire l'objet d'une explicitation claire et d'une analyse serrée consistant à examiner s'ils sont réellement en adéquation avec ce que l'on considère *in abstracto* comme un enseignement de qualité. L'exigence de transparence ne devrait pas se réduire à la nécessité de rendre publique la démarche évaluative et de la décrire d'une manière à ce point précise qu'elle soit répétable. Cette exigence devrait aller jusqu'à l'explicitation argumentée des valeurs au nom desquelles la démarche est mise en œuvre et de celles qui justifient le recours à telle ou telle modalité ou à tel ou tel critère.

4.4 Il n'est pas superflu de s'assurer que l'évaluation soit vraiment menée au service de l'amélioration

Dans les discours et comme pour faire « passer la pilule », l'évaluation est le plus souvent présentée comme étant essentiellement au service de l'amélioration de la qualité des pratiques, fonction régulatrice noble et

peu contestable. On évalue les acquis pour réguler les apprentissages, les enseignements pour les améliorer, les établissements pour mieux les piloter... Dans les faits, la fonction de contrôle prend bien souvent le dessus et l'évaluation participe davantage à la reddition de comptes, la sanction, la disciplinarisation des comportements, la gestion administrative, voire à la différenciation du financement.

Il n'est donc pas superflu de vérifier si l'évaluation est effectivement réalisée au bénéficie de l'amélioration des pratiques. Bien sûr, dans la réalité, ces deux types de fonctions sont bien souvent et inévitablement intriquées, comme le montrent les pratiques effectives d'évaluation des enseignements par les étudiants. L'équilibre est donc à chercher du côté d'un dosage acceptable entre le temps, l'énergie et les moyens consacrés à l'évaluation et ceux qui seront ultérieurement attribués aux procédures d'amélioration des pratiques, si vraiment tel est le but de l'évaluation.

L'importance accordée au suivi et à l'accompagnement post-évaluation constitue donc un bon indicateur de cet équilibre. Par exemple, on pourra vérifier si une politique d'évaluation des enseignements par les étudiants est réellement au service de l'amélioration des pratiques pédagogiques en réalisant le test proposé par Ricci (2009) : pour tout € investi dans le dispositif d'évaluation, un € est-il aussi investi dans l'amélioration en termes d'aide apportée aux enseignants à l'interprétation des résultats et à la recherche de solutions pédagogiques ? Dans le cas contraire, tous les doutes sont permis... Une exigence corollaire a été récemment proposée par Hadji (2012) : la frénésie évaluative dépasse sans doute les limites de la raison saine lorsqu'elle nuit à autrui plutôt que d'être à son service, ce qui pourrait, par exemple, être le cas d'une évaluation des enseignements standardisée et imposée provoquant un sentiment de découragement et d'incompétence auprès de jeunes enseignants.

4.5 L'évaluation suppose l'externalité, mais ne doit pas échapper aux acteurs

Malgré l'intérêt de l'auto-évaluation, une part d'externalité est sans doute nécessaire à toute démarche d'évaluation. Juge et partie, la personne évaluée n'est pas la mieux placée pour élaborer un avis suffisamment distancié sur son fonctionnement et sur ses productions. L'évaluation des publications de recherche par les pairs, celle des acquis des étudiants par les enseignants ou encore celle de l'efficacité d'un établissement ou d'une filière par un panel d'experts reposent sur ce sain principe.

Mais, en même temps, l'évaluation ne produit ses effets que si les acteurs ne se sentent pas dépossédés de la maîtrise du processus. Dans une situation idéale, l'évaluation est initiée par les acteurs eux-mêmes et réalisée en collaboration avec eux, depuis la définition de la démarche

et de ses étapes jusqu'à l'interprétation des résultats en passant par la détermination des critères. C'est le sens de la phase d'auto-évaluation pratiquée lors de l'évaluation institutionnelle des établissements, du portfolio de compétences élaboré par l'étudiant lors de l'évaluation de ses acquis ainsi que du dossier d'enseignement dans lequel les enseignants présentent, discutent et mettent en perspective les résultats qu'ils ont obtenus à l'occasion de l'évaluation de leurs enseignements par les étudiants.

Cette exigence de participation et de co-maîtrise des acteurs par rapport à l'ensemble du processus d'évaluation se heurte à la tendance actuelle vers la standardisation et l'automatisation des procédures (e.g. à partir de bases de données de publications pour l'évaluation de la recherche). La standardisation et l'automatisation participent en effet à l'externalisation du processus d'évaluation. Celui-ci finit par échapper aux acteurs et est mis en route, réalisé, contrôlé et même monnayé (dans le cas de l'évaluation de la recherche et des palmarès d'établissements) par des personnes externes au monde évalué.

En réalité, le caractère exogène de ces procédures constitue sans doute l'une des principales raisons de leur succès actuel : elles peuvent être appliquées par des personnes « incompétentes » dans le domaine, c'est-à-dire incapables de donner un avis sur la valeur intrinsèque de ce qui est censé être évalué. Ainsi, l'évaluation de la recherche sur la base d'indices de citation peut être mise en route par n'importe quel technicien maîtrisant la bibliométrie, indépendamment de ses compétences disciplinaires. Dans le même registre, une évaluation des enseignements par les étudiants réalisée sur la base de critères trop formels pourrait être menée en l'absence de prise en compte de la cohérence et de l'adéquation des contenus enseignés, comme l'a bien montré une expérimentation célèbre connue sous le nom de *Dr Fox experiment*. Un exposé farfelu et réalisé par un très bon acteur, ignorant du contenu traité, avait été jugé « clair et stimulant » par trois auditoires successifs... Il est donc illusoire de penser pouvoir mesurer l'excellence et la qualité sans comprendre quoi que ce soit au contenu de ce qui est à évaluer.

Cette exigence d'implication forte des acteurs plaide aussi en faveur d'une évaluation située et contextualisée, ici aussi à l'encontre d'une évaluation standardisée. Par exemple, l'évaluation d'un établissement devrait être fondée sur une analyse serrée de son contexte, de ses objectifs et de ses contraintes spécifiques. Il en va de même pour l'évaluation d'un programme, de laquelle, comme le montre le chapitre 7, on ne peut espérer recueillir des données pertinentes à son amélioration ultérieure que si l'on s'est posé les bonnes questions, c'est-à-dire les questions les plus adéquates à ses objectifs et à ses caractéristiques spécifiques.

Au final, c'est donc une image contrastée et paradoxale de la situation de l'évaluation dans l'enseignement supérieur qui se dégage du présent ouvrage. Certes, personne ne conteste sa nécessité : comment décerner des diplômes sans certifier des acquis ? Comment améliorer des programmes de formation sans en connaître les effets et sans recueillir l'avis de leurs bénéficiaires ? Comment faire progresser le savoir sans évaluer la pertinence de diffuser les résultats de la multitude des recherches menées ? De plus, des pratiques innovantes se sont développées et ont contribué à élargir le spectre de l'évaluation et à en enrichir les méthodes : la Valorisation des Acquis de l'Expérience, l'évaluation des enseignements par les étudiants, l'évaluation de programmes et de filières, l'évaluation par compétences... Mais la frénésie avec laquelle on évalue désormais tout et tout le temps inquiète. Par ailleurs, la standardisation, l'externalisation, l'automatisation et l'internationalisation croissantes des modalités d'évaluation sont dénoncées parce qu'elles sont mises en œuvre sans recul critique quant à leurs limites et parce qu'elles entraînent des effets dommageables importants.

Curieusement, un objet semble échapper miraculeusement au prurit évaluatif actuel : l'évaluation elle-même... Car après tout, il serait salutaire et cohérent d'appliquer aux différentes formes d'évaluation qui se sont déployées ces dernières années une démarche évaluative complète, ambitieuse et sans tabou. Ainsi, l'évaluation formative permet-elle vraiment aux étudiants d'améliorer leurs apprentissages et aux enseignants de jeter des ponts constructifs entre enseignement et évaluation ? L'évaluation des acquis de l'expérience contribue-t-elle à assouplir les parcours de formation et à faire accéder à l'enseignement supérieur des publics nouveaux ? Le défi que s'est donné l'enseignement supérieur d'aller jusqu'à évaluer des compétences, voire des compétences professionnelles ou du moins pré-professionnelles peut-il être considéré comme relevé ? Les technologies ont-elles contribué à renouveler durablement les pratiques d'évaluation ? L'évaluation des enseignements par les étudiants a-t-elle un impact sur l'amélioration des pratiques pédagogiques ? L'évaluation des programmes et des formations entraîne-t-elle leur amélioration continue ? L'évaluation des publications a-t-elle permis aux équipes et aux chercheurs de produire davantage de recherches de qualité ? Les établissements qui bénéficient d'une évaluation institutionnelle voient-ils augmenter l'efficacité de leur fonctionnement ? L'évaluation des filières de formation contribue-t-elle vraiment à leur pilotage ou celui-ci dépend-il plutôt de choix idéologiques ou de pressions environnementales ? Il est sans doute urgent de tenter de répondre de manière rigoureuse, outillée et empirique à ces questions, sous peine de devoir conclure que les zélateurs de l'évaluation sont étrangement les premiers à l'esquiver...